TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
Arvutiteaduse instituut
Infotehnoloogia eriala

Harri Kirik

Keelemudelipõhised parandused statistilises masintõlkes

Magistritöö (30 EAP)

Juhendaja: Mark Fišel

Autor: …… "….." mai 2010
Juhendaja: ……..….. "…." mai 2010

Lubada kaitsmisele
Professor Mare Koit ……..……. "….." mai 2010

TARTU 2010

Sisukord

Sissejuhatus 4
1 Statistiline masintõlge 5
1.1 Masintõlke olemus ja üldpõhimõtted 5
1.2 Sõnapõhine ja fraasipõhine statistiline masintõlge 6
1.3 Mürakanali mudel statistilises masintõlkes 8
1.4 Faktoriseeritud fraasipõhine statistiline masintõlge 10
1.5 Masintõlke kvaliteedi hindamine 12
1.6 Eksperimentide läbiviimiseks kasutatud rakendused 14
1.6.1 Giza++ 14
1.6.2 SRILM 15
1.6.3 Moses 15
1.6.4 TreeTagger 15
1.7 Eksperimentide läbiviimiseks loodud vahendid 16
2 Eksperimendid 17
2.1 Kasutatud tõlkesuunad ja korpused 17
2.2 Eksperimentide valdkond ja motivatsioon 18
2.3 Ettevalmistused 19
2.4 Läbiviidud eksperimendid 20
2.4.1 Baasmudeli loomine 20
2.4.2 Sõnavormidest koosnev teine faktor 22
2.4.3 Sõnaklasside alusel loodud teine faktor 25
2.4.4 Sõnaklasside ja esinemissageduste alusel loodud teine faktor 28
2.4.5 Ainult esinemissageduste alusel loodud teine faktor 31
2.4.6 Juhuslikkuse alusel loodud teine faktor 34
2.5 Tulemuste kontrollimine kasutades kolmandat tõlkesuunda 39
2.5.1 Baasmudel kolmandale tõlkesuunale 41
2.5.2 Esinemissageduste alusel loodud teine faktor (kolmas tõlkesuund) 41
2.5.3 Juhuslikkuse alusel loodud teine faktor (kolmas tõlkesuund) 42

2.6 Saadud tulemuste analüüs 45
Kokkuvõte 47
Language model based improvements in statistical machine translation 48
Kasutatud kirjandus 49
Lisad 51
Lisa 1. Loodud skriptid 52
Lisa 2. Sagedusnimekirjade näiteid 55

Sissejuhatus

Lingvistilise teadmuse kasutamine statistilise masintõlkes on populaarne valdkond, mida lähemalt uurib ka Tartu Ülikooli Matemaatika-informaatikateaduskonna masintõlke töögrupp. Käesolevas töös uuritakse ühte selle valdkonna võimalust, kus lisades tõlkimisel sihtkeele korpusesse lisainformatsiooni sõnaliikide näol, püütakse seda ära kasutada loomaks paremaid keelemudelite konfiguratsioone ja tõstmaks masintõlke väljundi kvaliteeti. Põhiideeks on lisaks tavalisele sõnavormidel loodavale primaarsele keelemudelile lisada veel sekundaarne, sõnaliikidel loodud mudel, mis aitaks tõlkimisel luua keeleliselt korrektsemaid laused.

Eksperimentide läbiviimiseks valiti algselt kaks tõlkesuunda, eesti-inglise ja inglise-prantsuse. Esimene tõlkesuund seetõttu, et vastav suund on masintõlke töögrupi üheks põhitõlkesuunaks, ja teine tõlkesuund seepärast, et kontrollida samade meetodite efektiivsust ka teistsuguseid keelenäiteid sisaldaval Europarl korpusel. Töö käigus võeti saadud tulemuste paremaks mõistmiseks kasutusele ka kolmas tõlkesuund, prantsuse keelest inglise keelde.

Käesolev töö koosneb kahest suuremast peatükist. Esimene peatükk tutvustab lugejale masintõlget, selle jaotumist, ning selgitab põhjalikumalt statistilise faktoriseeritud masintõlke tööpõhimõtteid. Samuti käsitletakse masintõlke hindamise probleemi ja tutvustatakse lühidalt eksperimentides kasutatavaid tarkvaralahendusi. Esimese peatüki põhiülesandeks on lugejale antud töö käigus tehtud eksperimentide paremaks mõistmiseks vajaliku taustinfo pakkumine, eeldades, et lugeja on üldise keeletehnoloogia valdkonnaga juba tuttav.

Teine peatükk on pühendatud töö käigus läbiviidud eksperimentidele. Selgitatakse tehtud katsete motivatsiooni, tehtud eeltöötlust ning kirjeldatakse lühidalt eksperimentide üldpõhimõtteid. Tuuakse ära saadud tulemused nii arendus kui ka testimiskorpusel, samuti arutletakse saadud tulemuste üle.

Tööga on kaasas ka lisad. Lisa 1 kirjeldab täpsemalt töö käigus loodud tähtsamaid skripte ning annab üldise nimekirja kõikidest autori poolt antud töö käigus kirjutatud abivahenditest. Lisa 2 toob näiteid eksperimentide jaoks koostatud sagedusnimekirjadest ning lisa 3 on CD arhiiv lisa 1 all kirjeldatud vahenditest.

1 Statistiline masintõlge

Alljärgnevas peatükis on lühidalt seletatud masintõlke ja selle statistilise variandi tööpõhimõtteid, samuti puudutatud tõlkehüpoteeside hindamise probleemi. Ära on toodud ka eksperimentide läbiviimisel tõlkimiseks kasutatud vahendid ja nende lühitutvustus. Käesoleva peatüki eesmärgiks ei ole masintõlke valdkonna sügavuti lahtiseletamine (selleks sobib hästi näiteks (Koehn, 2010)), vaid lugejale antud töö käigus tehtud eksperimentide paremaks mõistmiseks vajaliku taustinformatsiooni pakkumine.

1.1 Masintõlke olemus ja üldpõhimõtted

Masintõlge on keeletehnoloogia valdkond, mis tegeleb arvutisüsteemide abil tekstide tõlkimisega ühest keelest teise keelde. Valdkonnas tegeletakse nii tõlkimise läbiviimiseks vajalike algoritmide väljatöötamisega kui ka teiste kaasnevate (näiteks optimeerimise) probleemidega, mis tekivad antud algoritmide kasutamisel väljaspool arendus- ja teadustööd.

Üldiselt jagatakse masintõlge oma lähenemise põhimõtte järgi kaheks: reeglipõhine masintõlge ja statistiline masintõlge.

Reeglipõhises masintõlkes kasutab tõlkesüsteem tõlkimiseks juba olemasolevat, süsteemi autori poolt etteantud teadmiste hulka, milleks enamasti on erinevad sõnastikud ja tõlkereeglite kogumid. Seda kogumit süsteem enamasti ise ei tekita ega muuda, ainult kasutab. Seega on lihtsasti aimatav, et antud süsteem suudab tõlkida täpselt nii hea kvaliteediga, kui head on tema poolt kasutatavad andmed. Samuti on mõistetav, et nii erinevate keeltepaaride, kui isegi samade keeltepaaride aga erinevate tõlkesuundade (eesti-inglise või inglise-eesti), jaoks on vajalikud erinevaid reeglihulgad. See omadus teeb reeglipõhise tõlkesüsteemi laiendamise ja uute tõlkesuundade kasutuselevõtu kulukaks ning aeganõudvaks. Samuti vajab kirjeldatud viisil loodud tõlkesüsteemi arendamine enamasti keeletehnoloogias (ja vastavates keeltes) haritud ekspertide kasutamist, kes oleks üldse suutelised süsteemile vajaliku reeglitehulga looma.

Eelnevale lähenemisele mõnes mõttes vastandiks on statistiliste mudelite põhiline ehk statistiline masintõlge. Kui reeglipõhises masintõlkes tõlgitakse kasutades juba olemasolevaid spetsiifilisi lingvistilisi reegleid ja teadmisi antud kindla keelepaari kohta, siis statistilises masintõlkes kasutatakse reeglite asemel statistilisi mudeleid. Kuna loomulikud keeled on osutunud arvatust tunduvalt keerukamateks, siis on väga keeruline nende

modelleerimine ja inimesele lihtsalt jälgitaval kujul reeglitena kirjeldamine. Samas on tänu arvutite laiale levikule kogunenud suur hulk digitaliseeritud tekste, mis on mahult piisavalt suured, et neid kasutades oleks võimalik rakendada statistilisi meetodeid. Kasutades masinõppe printsiipe võimaldavad need statistilised meetodid luua automaatselt tõlkimisreeglite hulki, mis ei pruugi üldse olla inimesele arusaadavas vormis, kuid samas kirjeldavad keeli piisavalt hästi, et nende alusel oleks võimalik tõlkida uusi, treeningandmetes varem mittekohatud lauseid. Seega kahe keele vahelise statistilise masintõlke läbiviimiseks on vajalik ainult piisavalt suurte kahekeelsete paralleelkorpuste[1] olemasolu, mida saaks kasutada automaatselt tõlkimisreeglite hulga loomiseks.

Arusaadavalt annab statistilise masintõppe lähenemise kasutamine tõlkesüsteemile eelise paremas olemasolevate ressursside kasutuses, võimaldades hästi ära kasutata juba eksisteerivaid tekstihulki ja tõlkenäited ning suutes ühe ja sama süsteemiga treenida tõlkimiseks vajalikud mudelid suvaliste keelepaaride jaoks. Samuti pole statistiliste masintõlkesüsteemide koostamisel enam vajalik mõlemat keelt tundvate ekspertide kaasamine.

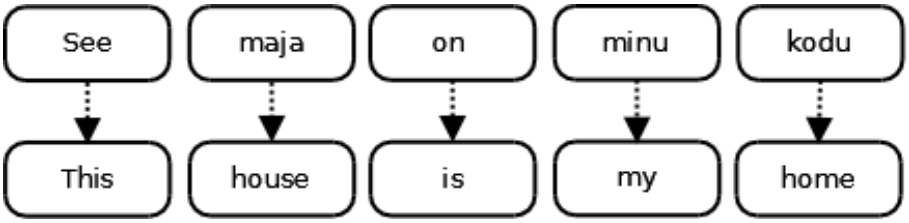
1.2 Sõnapõhine ja fraasipõhine statistiline masintõlge

Statistiline masintõlge jaguneb omakorda mitmeks erinevaks variandiks, millest siinkohal puudutame sõnapõhist ja fraasipõhist masintõlget.

Esimeseks statistilise masintõlke variandiks, mis loodi, oli sõnapõhine statistiline masintõlge. Nagu nimestki oletada võib, on antud lähenemisel vähimateks tõlgitavateks üksusteks sõnad. Kogu sõnade tõlkimiseks vajalik informatsioon, ehk kuidas tõlkida algkeelset sõna *f* sihtkeelseks sõnaks *e*, saadakse paralleelkorpuses sisalduvate näidete põhjal treenitud statistiliste mudelite abil. Tõlkimine ise käibki nn „sõna kaupa“, võimaldades kindla sõna tõlkimisel erinevaid tulemusi, näiteks et sõna algkeeles *f* tõlgitakse sihtkeelde *e* üheks sõnaks, mitmeks sõnaks või kaotatakse sihtkeelses lauses üldse. Samuti lubavad keerukamad mudelid uute sõnade tekitamist ja sõnade ümberjärjestamist sihtkeele lauses. Selle, milline variant eelkirjeldatud üks-mitmele suhtest (algkeele sõna võib tõlkida *0* .. *n* sihtkeele sõnaks) realiseerub, määrab tõlgitavad sõna viljakus (*fertility*) ja sihtkeele peal treenitud sõnade ümberjärjestamismudel (*reordering model*).

Joonis 1 illustreerib sõnapõhist masintõlget eesti keelsest inglise keelde.

[1] Paralleelkorpus on tekstikorpus, mis sisaldab seotud lausete paare: lause keeles A ja selle tõlge keeles B.

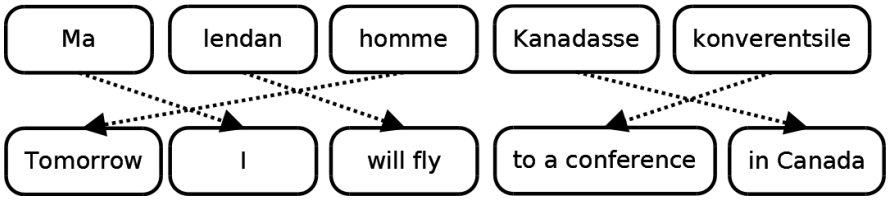


Joonis 1. Sõnapõhine statistiline masintõlge

Sõnapõhise masintõlke negatiivseks küljeks on, et viljakus töötab ainult ühes suunas. Seega ei ole antud süsteem võimeline tõlkima mitut algkeelset sõna üheks sihtkeelseks sõnaks. Üheks võimaluseks seda probleemi lahendada on fraasipõhise statistilise masintõlke kasutamine.

Fraasipõhises statistilises masintõlkes on tõlgitavad üksused sõnade asemel fraasid. Antud kontekstis ei pruugi fraas omada mingit lingvistilist põhjendust, vaid võib olla lihtsalt suvalise arvu üksteisele järgnevate sõnade järjend. Seda seepärast, et lausete jagamine fraasideks tõlgitavas tekstis toimub tavaliselt statistiliste andmete alusel, arvestamata vaadeldava keele süntaksit. On näidatud (Koehn, Och, & Marcu, 2003), et selline statistikapõhine fraasistruktuuri loomine annab positiivseid tulemusi, kuna paralleelkorpustes on peale süntaktiliselt motiveeritud fraaside ka tihti väga sagedasti esinevaid sõnajärjendeid, mida mõttekas fraasidena käsitleda, ja mitte piirata fraasipõhist statistilist masintõlget kasutama ainult süntaktiliselt motiveeritud fraase.

Fraasipõhise lähenemise põhieeliseks on see, et enam ei tõlgita sõnu sõnadeks, vaid tõlgitakse tervete fraasida kaupa ning seepärast võetakse paremini arvesse sõnade lokaalset konteksti, kuna sageli sõltub tõlgitava sõna tähendus mingil kindlal esinemisel just antud sõnale eelnevast või järgnevast sõnast. Joonis 2 illustreerib fraasipõhist masintõlget eesti keelsest inglise keelde.



Joonis 2. Fraasipõhine statistiline masintõlge

1.3 Mürakanali mudel statistilises masintõlkes

Tõlkimise läbiviimiseks statistilises masintõlkes kasutatakse kolme põhikomponenti, mis omavahel kombineeritult moodustavad kõnetuvastuses paljukasutatud, algselt Claude Shannon'i informatsioonteooriast (*Information theory*) (Shannon, 1948) pärineva mürakanali mudeli (*Noisy channel model*). Mainitud kolm komponenti on järgnevad:

1. Tõlkemudel (*translation model*)
2. Keelemudel (*language model*)
3. Dekooder (*decoder*)

Eelneva illustreerimiseks oletame, et soovime tõlkida eestikeelset lauset *f* ingliskeelseks lauseks (ehk tõlkehüpoteesiks) *e*. Sellisel juhul võimaldab mürakanali mudel anda võimalikele tõlkehüpoteesidele hinnangu, kui (statistiliselt) sarnane on vaadeldav hüpotees eelnevalt nähtud lausetele treeningandmetes. Selline hinnang ê hüpoteesile *e* on kirjeldatud kui:

$$\hat{e} = argmax_e\, p(e) * p(f|e).$$

Eelnevas valemis on olemas kõik mainitud kolm mürakanali mudeli põhikomponenti:

- $p(e)$ ehk keelemudel. Saab sisendiks genereeritud ingliskeelse tõlkehüpoteesi *e*, väljastab antud ingliskeelse hüpoteesi *e* tõenäosuse keelelise korrektsuse seisukohalt. Täpsemalt mudel väljastab treeningandmete põhjal arvutatud tõenäosuse, mis näitab kui suure tõenäosusega hüpotees *e* kuulub treeningandmetes kohatud korrektsete ingliskeelsete lausete hulka.

  Üks laialt levinud keelemudeli realisatsioon, mida ka antud töös kasutatavad programmid rakendavad, kasutab hüpoteesile *e* hinnangu arvutamisel n-gramme, omistades suurema arvu n-grammidega arvutatud tõenäosusele suurema kaalu.

  Näiteks koosnegu tõlkehüpotees *e* sõnadest $S_0, S_1, S_2, .., S_i, .., S_n$. Siis 3-grammilist hinnangukomponenti hüpoteesi *e* sõnale S_i saab leida järgneva valemi abil:

$$p(e)_3 = \prod_i p(S_i|S_{i-1} * S_{i-2})$$

- $p(f|e)$ ehk tõlkemudel. Saab sisendiks eestikeelse lähtelause *f* ja genereeritud ingliskeelse tõlkehüpoteesi *e,* ning väljastab tõlke $f \Longrightarrow e$ tõenäosuse. Mida paremini vastavad laused *e* ja *f* treeningandmetes nähtud tõlkepaaride põhjal üksteisele, seda suurem on tõlkemudeli väljastatud tõenäosus. Fraasipõhises masintõlkes koosneb tõlkemudel kahest alamkomponendist, leksikalisest vastavuste mudelist $p_{lex}(f|e)$ ja fraaside järjekorramudelist $p_{dist}(f|e)$. Leksikaline vastavuste mudel hindab seda, kui täpselt vastavad lähtelause *f* ja genereeritud hüpotees *e*. Fraaside järjekorramudel tegeleb tõlkel fraaside järjestuse ümberpaigutuse hindamisega. Seega fraasipõhise masintõlke tõlkemudelit illustreerib valem:

$$p(f|e) = p_{lex}(f|e) * p_{dist}(f|e)$$

- $argmax_e$ ehk dekooder. Leiab eestikeelse lause *f* jaoks suurima tõenäosusega ingliskeelse tõlkehüpoteesi *e*. Seega dekooder „liigub“ suures otsinguruumist, kasutades otsingutulemuste hindamiseks keele- ja tõlkemudelit, eesmärgiga leida optimaalne tõlge.

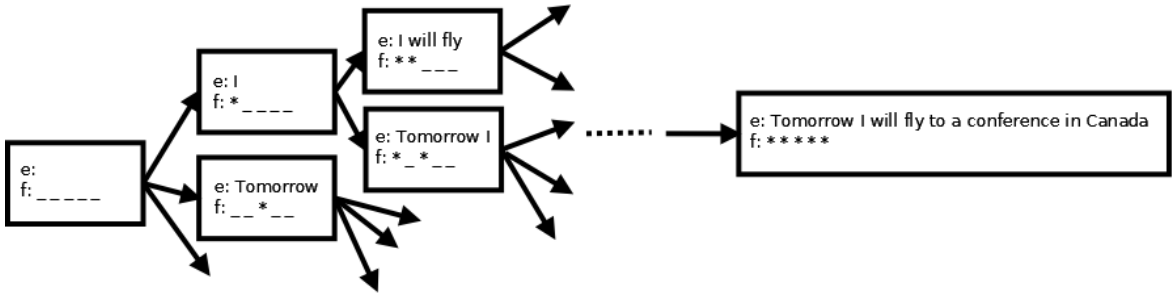
Võttes kasutusele eelnevalt kirjeldatud valemi, seisneb algkeelsele lausele *f* sihtkeele lause ehk korrektse tõlke *e* genereerimine suures otsinguruumis parima väljundfraasi konstrueerimises. Kuna erinevatest loomuliku keele fraasidest koosnev otsinguruum on mittelõplik, siis ei ole enamasti võimalik leida parimat tõlget, vaid läbi erinevate optimeerimistehnikate püütakse leida mõni piisavalt hea tõlge.

Tõlkehüpoteesi otsimine dekoodri poolt toimub ühe võimaliku variandi (Koehn, 2004) kohaselt järgnevalt: kasutades otsimisrinnet (*beam search*), genereeritakse tõlge liikudes tõlkehüpoteesi pidi vasakult paremale. Algseisuks on olukord, kus ühtegi algkeele sõna pole sihtkeelde tõlgitud. Iga uus seis tekib, kui eelnevat seisu täiendatakse mingi tõlgitud fraasiga, mis katab osa algkeelsest lausest. Iga uut seisu hinnatakse kaaluga, mille väärtuseks on eelmise seisu kaal korrutatuna tõke- ja keelemudeli poolt lisatud fraasile antud kaaludega. Samuti on igas seisus teada mõningad andmed eelmise seisu ja käesolevasse seisu jõudmiseks tehtud sammu kohta.

Hüpoteesi tõlkimisel saavutatakse lõplik seis, kui on tõlkega kaetud kõik algkeelsed sõnad. Kuna enamasti genereeritakse dekoodri töö tulemusena rohkem kui üks võimalik tõlke-

hüpotees, siis väljastatavaks tõlkeks valitakse tõlge, mis on kõige väiksema kaaluga. Samas kasutatakse ka teisi võimalusi, näiteks väljastatakse *n* parimat tõlget ja rakendatakse mingit muud vahendit nende seast parima valimiseks.

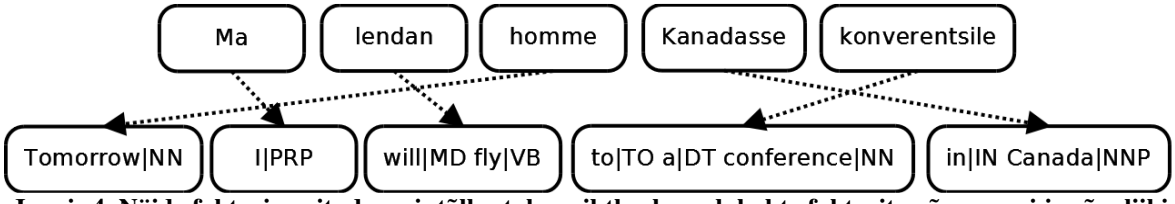
Joonis 3 illustreerib dekoodrit tõlkimas eestikeelselt lauset „*Ma lendan homme Kanadasse konverentsile*“ inglise keelde („*Tomorrow I will fly to a conference in Canada*“). Joonisel tähistab märgend *e* sihtkeelt, märgend *f* algkeelt, tärnid on juba tõlgitud sõnad, kriipsud veel tõlkimata sõnad.



Joonis 3. Dekoodri tööpõhimõtte illustratsioon

1.4 Faktoriseeritud fraasipõhine statistiline masintõlge

Faktoriseeritud fraasipõhine masintõlge erineb tavalisest fraasipõhisest masintõlkest selle poolest, et saab sisendiks sõnedest koosneva lause asemel vektoritest koosneva lause, kus vektorite elementideks võib olla suvaline informatsioon, sealhulgas näiteks morfoloogilised või süntaktilised andmed. Joonis 4 illustreerib faktoriseeritud masintõlke sisendandmeid, kus algkeelel on üks faktor – sõnavorm, sihtkeelel kaks (püstkriipsuga eraldatud) faktorit – sõnavorm ja sõnaliik.

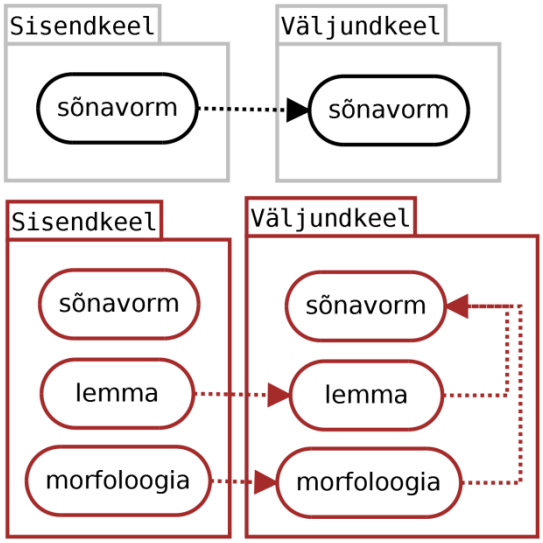


Joonis 4. Näide faktoriseeritud masintõlkest, kus sihtkeel omab kahte faktorit - sõnavormi ja sõnaliiki

Faktorite kasutamine annab võimaluse tänu sisendandmetes leiduvale lisainformatsiooni kättesaadavusele treenimisprotsessis luua paremaid keele- ja tõlkemudeleid.

Tõlkemudeli puhul on sagedasti esitatud näiteks järgnev võimalus: selle asemel, et õppida tõlkima otse algkeele *f* sõnavormist sihtkeele *e* sõnavormi, võib morfoloogiliselt rikaste keelte puhul olla kasulikum tõlkida sõna lemma ja morfoloogiline informatsioon eraldi,

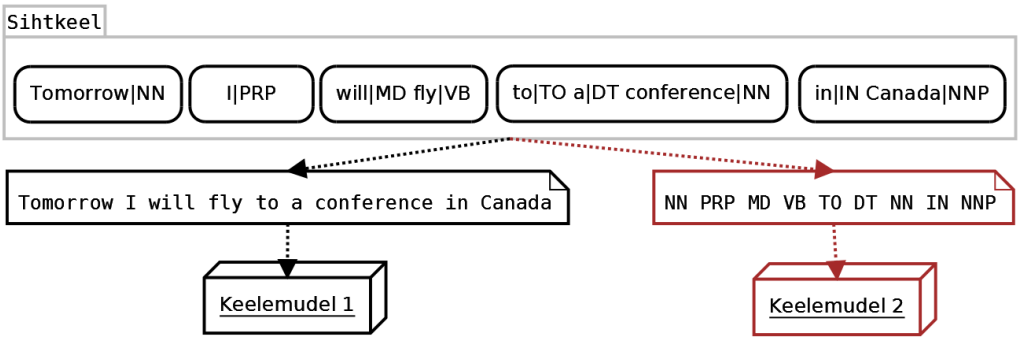
ning nende andmete põhjal sihtkeelde tõlgitud lemma õigesse morfoloogilise vormi panna. Allolev joonis 5 illustreerib kirjeldatud võimalust:



Joonis 5. Faktorites leiduva informatsiooni kasutamine tõlkemudelis

Lisaks eelnevale variandile pakub tõlkemudeli genereerimine veel mitmeid erinevaid võimalusi lisainformatsiooni kasutamiseks mudeli treenimisel, näiteks lemmade kasutamist parema sõnajoonduse ja -järjestuse saamiseks (Kirik, 2008) või lemmade kasutamist alternatiivse tõlkena sõnavormide tõlkele (Kirik & Fishel, 2008; Fishel & Kirik, 2010).

Keelemudeli puhul saab faktorites leiduvat lisainformatsiooni kasutada näiteks lisa-keelemudeli loomiseks. Käesolevas töös me uurimegi sellist võimalust, kus kasutame tõlkimisel mitte ühte, vaid kahte erinevat keelemudelit. Esimene keelemudel luuakse sihtkeele sõnavormide pealt nagu masintõlkel enamasti tavaks. Aga teine keelemudel luuakse sihtkeele sõnaliikide pealt, mis sisaldavad küll vähem informatsiooni kui sõnavormid, kuid samas võimaldavad saada rohkem ja sagedasemaid treeningandmeid (sõnavormide hulk on loomulikes keeltes üldiselt tehnilisest seisukohast mittelõplik, aga sõnaliike on lõplik hulk) ning luua suurema keelemudeli. Tõlkimisel kasutab dekooder mõlemat keelemudelit korraga, kombineerides nende antavaid hinnanguid (vastavalt keelemudelitele antud kaaludele). Joonis 6 illustreerib faktorite alusel erinevate keelemudelite loomist:



Joonis 6. Faktorites sisalduva informatsiooni põhjal erinevate keelemudelite loomine

1.5 Masintõlke kvaliteedi hindamine

Kõikide loomuliku keele töötluse ülesannete (ka masinõppe probleemide) puhul on tähtis omada viisi, kuidas olemasolevat mudelit või mudelile tehtud parandusi hinnata.

Kui mõelda tõlke kvaliteedi hindamisest, tuleb kindlasti kohe pähe olukord, kus nii alg kui sihtkeeles pädev tõlk masintõlkesüsteemi poolt loodud tõlkeid (edaspidi „tõlkehüpoteese“) vaatab ning nende kvaliteedile oma hinnangu annab. Selline lahendus annab küll enamasti (eriti mitme eksperdi koos kasutamise puhul) tõlkehüpoteesile väga täpse hinnangu, kuid on oma olemuselt aeganõudev ja kallis ning pole mõeldav inkrementaalsete muutuste kiireks hindamiseks. Aga just seda viimast omadust on masintõlke arendajatel oma töös enamasti kõige rohkem vaja.

Teiseks levinud hindamisviisiks on mingi arvuti poolt arvutatava meetrika kasutamine. Üks selline meetod on näiteks BLEU (*Bilingual Evaluation Understudy*) (Papineni, Roukos, Ward, & Zhu, 2002). BLEU meetrika seisneb põhimõttel, et mida lähedasem on tõlkehüpotees professionaalse inimtõlgi loodud tõlkele (edaspidi „etalontõlge“), seda parem antud masintõlkesüsteem tõenäoliselt on. BLEU kasutamine on populaarne seetõttu, et ta on üks esimesi automaatseid hindamisviise, mis enamasti omab kõrget korrelatsiooni inimhindajate poolt tõlgetele antud punktidega. Samas on tähtis meeles pidada, et kuna antud meetrika hindab ainult tõlkehüpoteesi sarnasust etaloniga, siis ei arvesta see meetrika otseselt tõlke grammatilist või sisulist korrektsust. Samuti on näidatud (Callison-Burch, Osborne, & Koehn, 2006), et BLEU kipub oma implementatsioonist tulenevalt eelistama statistiliste masintõlkesüsteemide tõlkeid teiste, mittestatistilise süsteemide tõlgetele. Selle probleemi üheks põhjuseks peetakse BLEU ja statistilise masintõlke tööpõhimõtete sarnasust ja n-grammide kasutamist neis mõlemas. Kuid kuna enamasti kasutatakse BLEU

meetrikat just ühele ja samale süsteemile tehtavate paranduste hindamiseks (nagu ka käesolevas töös), siis pole antud probleem meetrika kasutamisel väga suureks takistuseks.
BLEU meetrika paremaks illustreerimiseks olgu meil tõlkehüpotees e_h ja etalontõlge e nagu kujutab tabel 1:

Tabel 1. Etalontõlge (e) ja tõlkehüpotees (e_h)

$\boldsymbol{e}$:	See on Mari kodu.
$\boldsymbol{e_h}$:	See on Mari maja.

BLEU arvutamiseks tuleb kõigepealt leida tõlkehüpoteesi e_h n-grammiline täpsuse hinnang (*ngram precision*), mis on defineeritud järgnevalt:

$$täpsus_n\ e_h = \frac{m}{p},$$

kus p on lause e_h n-grammide koguarv ja m lauses e_h esinenud selliste n-grammide arv, mis esinesid ka etalontõlkes e.
Tabelis 1 sisalduva näite korral on näiteks unigramm ja trigramm täpsused järgnevad:

$$täpsus_1\ e_h = \frac{3}{4} = 0{,}75$$

$$täpsus_3\ e_h = \frac{1}{2} = 0{,}25$$

Nüüd, kus meil on defineeritud n-grammilise täpsuse leidmiseks vajalik valem, saab defineerida BLEU meetrika:

$$\text{BLEU} = \min\left(1, \frac{pikkus_{e_h}}{pikkus_e}\right) * \prod_{i=1}^{n} täpsus_i{}^{\lambda_i}$$

BLEU valemi esimene komponent:

$$\min\left(1, \frac{pikkus_{e_h}}{pikkus_e}\right)$$

on nn lühidustrahv (*brevity-penalty*), mis penaltiseerib liiga lühikesi tõlkeid, ehk tõlkeid, kus on sõnu vähem kui etalontõlkes. See kompenseerib olukorra, kus tõlkides ainult neid sõnu, mille tõlkeid kindlalt teatakse, saadakse väljundi väga kõrge täpsuse hinnang, samas aga praktiliselt tähenduseta tõlge.

Korrutises

$$\prod_{i=1}^{n} täpsus_i^{\lambda_i}$$

määrab muutuja *n* suurima n-grammi suuruse, mille jaoks täpsus arvutatakse. Tüüpiliselt on selle vaikimisi väärtuseks $n = 4$ (Koehn, 2010, lk 226-228), millisel juhul võib meetrikat nimetada ka kui BLEU-4.

BLEU üks häid omadusi on see, et see meetrika suudab kasutada ka mitut etalonlauset korraga. Sellisel juhul on suurem tõenäosus, et BLEU annab adekvaatse tulemuse ka lausete puhul, mis on küll korrektsed tõlked, aga pole sõnakasutuse poolest sarnased ühe kindla etalonlausega.

1.6 Eksperimentide läbiviimiseks kasutatud rakendused

1.6.1 Giza++

Giza++[2] (Och & Hermann, 2003) on 1999. aastal Johns-Hopkins ülikoolis loodud statistilise masintõlke tõlkemudelite treenimise rakendus. Giza++ rakendust kasutatakse treenimaks IBM mudeleid (*IBM translation models*) (Brown, Pietra, Pietra, & Mercer, 1993) ja Markovi peitmudelil (HMM) põhinevaid sõnajoondusmudeleid. Giza++ loodud mudelid sobivad sisendiks antud töös kasutatud Moses masintõlkesüsteemile. Kuid kuna Giza++ poolt loodud mudelid on sõnapõhised tõlkemudelid, tuleb need eelnevalt Moses süsteemis kasutamiseks muuta fraasipõhisteks mudeliteks. Selle muutuse läbiviimiseks on Moses süsteemil kaasas vastav skript.

[2] Giza++ on allalaetav lehelt: http://fjoch.com/GIZA++.html

1.6.2 SRILM

SRILM[3] (Stolcke, 2002) on Andreas Stolcke poolt loodud kõnetuvastuse ja statistilise masintõlke jaoks keelemudelite treenimise rakendus. SRILM võimaldab luua Moses masintõlkesüsteemi poolt kasutatavaid sihtkeele keelemudeleid.

1.6.3 Moses

Moses[4] (Koehn, et al., 2006) on vabavaraline lahtise lähtekoodiga (LGPL) statistilise masintõlkesüsteemi dekooder, millel jaoks on olemas ka skriptid ja tööriistad tõlke- ning keelemudeli treenimiseks. Moses masintõlkesüsteemi on võimalik rakendada kahe suvalise keelepaari jaoks, eelduseks on ainult piisavalt suure paralleelkorpuse olemasolu.
Statistiline masintõlge süsteemiga Moses toimub kolmes etapis:

1. Genereeritakse sihtkeele keelemudel (*language model*), kasutades näiteks rakendust SRILM (*SRI Language Modeling Toolkit*), või mõnda alternatiivset keelemudelite loomise vahendit.
2. Luuakse keeltevaheline tõlkemudel (*translation model*) kasutades kaasasolevat vahendit GIZA++ sõnapõhise mudeli treenimiseks, ja skripti *train-factored-phrase-model.perl* sõnapõhisest mudelist fraasipõhise mudeli saamiseks.
3. Kasutatakse Moses dekoodrit teksti tõlkimiseks lähetekeelest sihtkeelde. Saadakse nn tõlkehüpotees, mida on võimalik võrrelda olemasoleva etalontõlkega.

Moses masintõlkesüsteem töötab nii Linux kui ka Windows operatsioonisüsteemidega, suudab hästi ära kasutada mitmeprotsessorilisi keskkondi ning toime tulla suurte, RAM mällu mittemahtuvate tõlke- ja keelemudelitega. Moses masintõlkesüsteemi arendus on toetatud EuroMatrix[5] projekti ja Euroopa Komisjoni poolt.

1.6.4 TreeTagger

TreeTagger[6] (Schmid, 1994) on Stuttgarti ülikoolis Helmut Schmid'i poolt loodud teksti sõnaliikide ja lemmadega märgendamise tööriist. TreeTagger töötab kasutades binaarseid otsustuspuid ja on treenitud ning katsetatud järgnevatel keeltel: saksa, inglise, prantsuse,

[3] SRILM on allalaetav lehelt: http://www-speech.sri.com/projects/srilm/download.html
[4] Moses on allalaetav lehelt: http://sourceforge.net/projects/mosesdecoder/files/
[5] Lisainfo projekti kohta lehel: http://www.euromatrix.net/
[6] TreeTagger on allalaetav lehelt: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

itaalia, hollandi, hispaania, bulgaaria, vene, kreeka portugali ja hiina keel. Vajadusel on võimalik märgendajat treenida mõne teise keele jaoks, tarvilik on omada vastava keele leksikoni ja käsitsi märgendatud treeningkorpust.

1.7 Eksperimentide läbiviimiseks loodud vahendid

Käesolevas töös tehtud eksperimentide läbiviimiseks ja andmete ettevalmistamiseks sai kirjutatud suur hulk erinevaid Python programmeerimiskeele programme ja samuti Linuxi käsureaskripte. Erinevalt eelnevast tööst (Kirik, 2008) sai seekord valitud modulaarne lähenemine, kus kõik loodud skriptid on enamasti lühikesed ja mingi kindla ülesande täitmiseks, samuti vähese vaevaga omavahel kombineeritavad ning täiendatavad. Lisaks sai abivahendite loomise põhikeeleks valitud Java asemel Python, kuna selle puhul puudub kompileerimisvajadus ning väikeste muutuste tegemine on seetõttu tunduvalt kiirem. Peale selle suurenes ka Python keele kasutuselevõtu tõttu programmide taaskasutuse võimalus, kuna erinevalt Java keelest olid töögrupi teised liikmed sellega juba tuttavad.
Kõik loodud skriptid on kirjeldatud peatüki lisa 1 all ning kaasatud ka tööle lisatud CD peal olevas arhiivis.

2 Eksperimendid

Alljärgnev peatükk on pühendatud läbiviidud eksperimentidele. Alustuseks selgitatakse läbiviidud katsete motivatsiooni, tehtud korpuste eeltöötlust ja ettevalmistusi. Samuti tuuakse ära saadud tulemused nii arendus- kui ka testimiskorpusel. Peatükk lõpeb tulemuste analüüsi ja järeldustega.

2.1 Kasutatud tõlkesuunad ja korpused

Käesolevas töös kasutatakse statistilist fraasipõhist faktoriseeritud masintõlget ja vaadeldakse erinevaid võimalusi faktoriseeritud korpuse abil lisakeelemudelite loomiseks kasutades kahte põhitõlkesuunda ja kahte erinevat korpust.

Tõlkesuundadeks on valitud eesti-inglise suund (järgnevalt nimetatud ka kui „esimene" tõlkesuund), kuna tõlge sellel suunal on Tartu Ülikooli Matemaatika-informaatikateaduskonna masintõlke töögrupi üheks tähtsamaks tööülesandeks, ning inglise-prantsuse tõlkesuund (ka „teine" tõlkesuund), kuna see annab võimaluse kontrollida samade meetodite tulemuslikkust ka teistsugust korpust ja keelepaari kasutades.

Korpustest sai valitud eesti-inglise tõlkesuuna jaoks JRC-Acquis (Steinberger, et al., 2006) seadustekstide korpus ja inglise-prantsuse tõlkesuuna jaoks Europarl (Koehn, 2005) tekstide korpus.

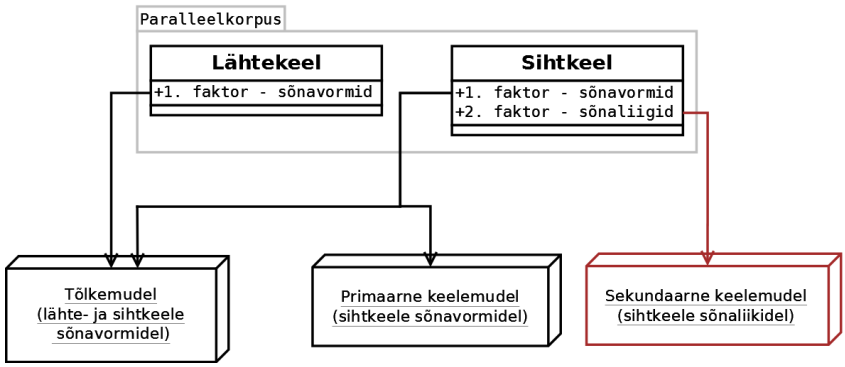
JRC-Acquis korpuse valiku kriteeriumiks oli, et see on üks vähestest korpustest, mis on masintõlke ülesande jaoks piisavalt mahukas ja samas sisaldab eesti-inglise paralleelkorpust. Antud korpus kätkeb endas valikut Euroopa Liidu liikmesriikide kohustusi ja õigusi kirjeldavast seadustekstide kogumist[7], mis vastavalt oma eesmärgile peab olema saadaval kõikide liikmesriikide keeltes.

Inglise-prantsuse suuna jaoks sai Europarl korpus valitud seetõttu, et oleks võimalik samasuguseid eksperimente läbi viia keelekasutuse küljelt erinevate korpuste peal. Nimelt sisaldab Europarl endas küll seadusloomealaseid tekste, aga erinevalt JRC-Acqis korpusest on need suulise kõne transkriptsioonid, mis sisaldavad endas Euroopa Parlamendi istungite poliitilisi väitlusi.

[7] Nimetatud seadustekstide kogumi nimi on: „*Acquis Communautaire*".

2.2 Eksperimentide valdkond ja motivatsioon

Praktiliselt alati kasutatakse masintõlkes ainult ühte keelemudelit, mis on loodud sihtkeele korpuse sõnavormide peal, ning sageli on see keelemudel loodud masintõlkerakenduse poolt pakutud vaikimisi suurusega. Käesoleva töö üldiseks põhimõtteks oli proovida kasutada morfoloogilist lisainfot sõnaliikide näol, et luua paremini sihtkeelt modelleerivaid keelemudeleid kasutades sama hulka treeninglauseid. Märgendades treeningkorpuse sihtkeele kõik sõnad sõnaliikidega (ehk lisades sõnaliiki sisaldava faktori korpusesse), sai luua primaarse keelemudeli kõrvale sekundaarse keelemudeli sõnaliikide pealt, nagu illustreerib joonis 7. Kuna sõnaliikide arv keeles on lõplik (ja tunduvalt väiksem kui sõnavormide arv), siis peaks selle faktori pealt keelemudelit luues saama tunduvalt rohkem korduvaid näiteid korjata kui sõnavormide pealt. Ning kombineerides väiksemat keelemudelit sõnavormidel ja suuremat keelemudelit sõnaliikidel, peaks olema võimalik saada sihtkeeles korrektsemaid tõlkeid. Selle oletuse kontrollimises seisnebki antud töö põhieesmärk.



Joonis 7. Sekundaarse keelemudeli loomine sõnaliikide põhjal

Käesoleva töö eksperimente alguses tehti kindlaks mõlema valitud keelepaari jaoks optimaalse suurusega primaarne keelemudel, mis on kirjeldatud esimese eksperimendina alampeatüki 2.4 all. Optimaalsed keelemudelid loeti mõlema tõlkesuuna jaoks nn baassüsteemideks (*baseline systems*), mida aluseks võttes hinnatakse järgnevate eksperimentide edukust. Järgmiseks asuti tegema katsetusi sekundaarsete keelemudelite loomisega (ehk siis kasutades tõlkimisel kahte keelemudelit), eesmärgiga leida selline loomise strateegia, mis parandaks baassüsteemidega saavutatud tulemusi. Need eksperimendid moodustavadki ülejäänud osa alampeatükist 2.4.

2.3 Ettevalmistused

Eksperimentide läbiviimiseks kasutatud korpused olid eelnevalt statistilises masintõlkes kasutamiseks eraldi töödeldud. Samuti oli tehtud mõlemast (JRC-Acquis ja Europarl) korpusest ka selline versioon, kus sihtkeele korpuses oli kaks faktorit: sõnavorm ja sõnaliik. Järgnevalt on kirjeldatud töötlust, mis viidi läbi kasutatud korpustel:

1. Esialgsest paralleelkorpusest moodustati kaks erinevat keelt sisaldavat korpust
2. Eemaldati XML märgendid
3. Kogu tekst mõlemas korpuses muudeti väiketähtedeks
4. Kirjavahemärgid eraldati: „see, muuseas, on test!" ⇒ "see , muuseas , on test !"
5. Asendati sümbolid, mis esinesid korpuse töötlemiseks kasutatud programmide siseloogikas
6. Eemaldati laused, mis olid pikemad kui 100 sõna ja samuti laused, kus lause ja antud lause tõlke vaheline sõnade arv erines rohkem kui üheksa sõna võrra
7. Eemaldati kõik laused, milles polnud vähemalt nelja järjestikkust tähte
8. Korpused jagati treenimiskorpuseks, arenduskorpuseks ja testimiskorpuseks
    a. arendus- ja testimiskorpusest eemaldati laused, mis esinesid ka treeningkorpuses
    b. testimiskorpusest eemaldati laused, mis esinesid arenduskorpuses

- Treenimiskorpust kasutati Moses süsteemi treenimiseks antud keelepaari jaoks, arenduskorpust eksperimentide läbiviimiseks ja tõlkemudelite hindamiseks. Saadud hinnanguid võimaldas kontrollida eraldi testimiskorpuse kasutamine.

Tabel 2 toob ära peale eeltöötlust saadud korpuste suurused nii JRC-Acquis kui ka Europarl korpuse puhul.

Tabel 2. Korpuste suurused peale eeltöötlust

JRC-Acquis korpus (eesti-inglise)	**Europarl korpus** (inglise-prantsuse)

• treenimiskorpus 1 088 389 lauset
  ○ eesti osa – 20 180 623 sõna
  ○ inglise osa – 27 911 130 sõna
• arenduskorpus 2500 lauset
• testimiskorpus 2500 lauset
	• treenimiskorpus 1 277 860 lauset
	  ○ inglise osa – 35 425 900 sõna
	  ○ prantsuse osa – 40 803 011 sõna
	• arenduskorpus 2500 lauset
	• testimiskorpus 2500 lauset

2.4 Läbiviidud eksperimendid

Järgnevalt on ära toodud kõik läbiviidud eksperimendid. Iga eksperimendi alguses on seletatud antud katse motivatsiooni ja üldsõnaliselt ära toodud katse etapid. Jälgimise lihtsustamiseks on iga eksperimendi juures ära toodud nii arenduskorpuse (joonistel märgitud kui „dev“) kui ka testimiskorpuse (joonistel märgitud kui „test“) BLEU skoorid. Samas tuleb rõhutada, et eksperimentide läbiviimisel kasutati arenduskorpust eksperimentide tulemuste esmaseks hindamiseks ja võrdluseks teise eksperimentidega ning testimiskorpuse tulemusi kasutati alles hilisemal etapil kontrollimaks, kas need toetavad arenduskorpusel saadud tulemusi.

Lisaks eksperimendi enda tulemustele võib iga eksperimendi juures olla ka veel võrdlus eelneva (selle tõlkesuuna parima) eksperimendi või baassüsteemi eksperimendi tulemustega.

Eksperimentide läbiviimisel kasutati üldjoones järgnevaid samme:

1. Märgendati korpuse sihtkeelne osa TreeTagger märgendaja abil sõnaliikidega ning märgendust töödeldi vastava eksperimendi jaoks loodud skripti(de)ga (antud sammu ei tehtud baaseksperimentide puhul)
2. Loodi vajalike suurustega keelemudelid kasutades SRILM rakendust
3. Loodi tõlkemudel kasutades GIZA++ rakendust
4. Loodi tõlkehüpoteesid Moses rakendusega, kasutades erineva suurusega (vastavalt kas siis primaarseid või sekundaarseid) keelemudeleid
5. Arvutati saadud tõlkehüpoteesidele BLEU meetrika skoorid

2.4.1 Baasmudeli loomine

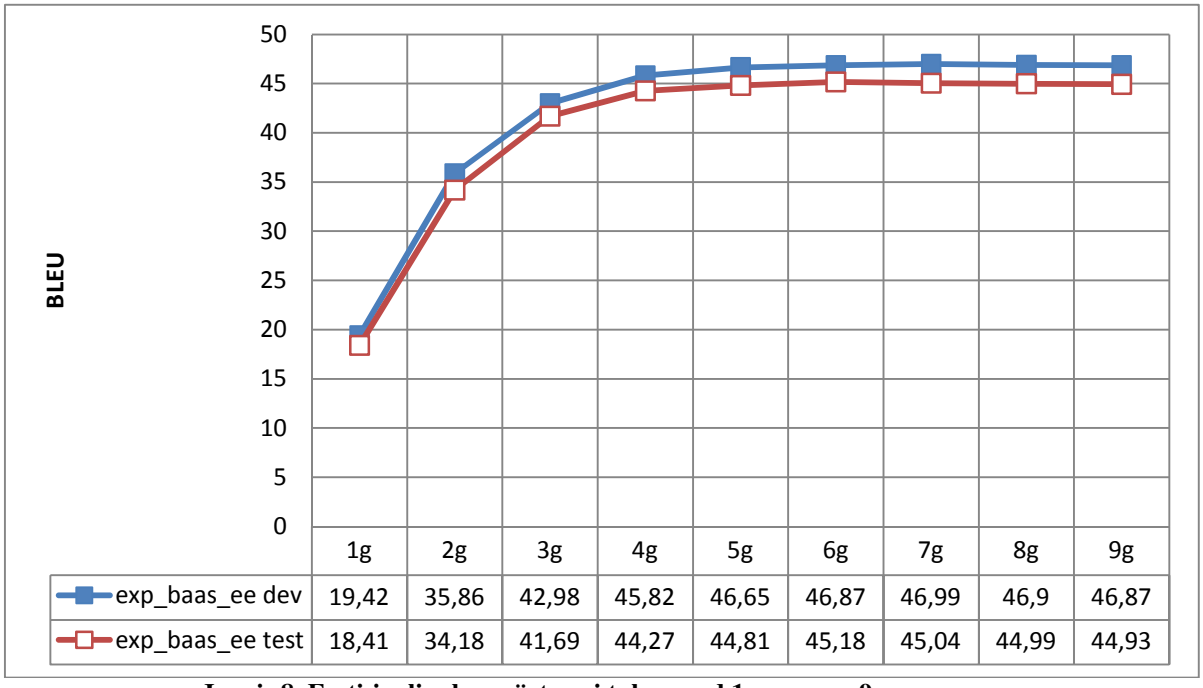
Enne põhieksperimentide läbiviimist oli ülesandeks paika panna nn baassüsteemide BLEU skoorid, mille põhjal oleks võimalik hinnata järgnevate eksperimentide edukust. Kuigi grupisiseselt on enamasti basssüsteemides kasutatud 3-gramm keelemudelit, siis kuna antud töö eesmärgiks on keelemudelipõhiste paranduste uurimine, sai katsetatud mõlema keelepaari korral suuremat hulka erineva suurusega keelemudeleid: eesti-inglise puhul unigramm mudelist kuni 9-gramm mudelini, ja inglise-prantsuse puhul unigramm mudelist kuni 7-gramm mudelini. Keelemudelite maksimaalsed suurused sai valitud sellised see-

tõttu, et meile kasutada antud riistvara[8] ei suutnud antud korpuste korral suuremate mudelite loomist läbi viia.

Teised parameetrid peale keelemudeli suuruste olid baassüsteemis konstantsed, sai kasutatud rakenduste vaikeväärtusi. Baassüsteemide loomine oli mõlema keelepaari korral sarnane, erinesid ainult etteantud treeningkorpused.

Käesolevas töös on baassüsteemi eksperimendi tähiseks *exp_baas_ee* (eesti-inglise korpuse puhul) või *exp_baas_ef* (inglise-prantsuse korpuse puhul).

Joonis 8 toob ära eesti-inglise korpusel saadud tulemused. Nagu näha, siis kuigi alguses tõuseb BLEU skoor iga suurema mudeli puhul märgatavalt, siis vahemikus 6-gramm kuni 9-gramm skoori tõus stabiliseerib, ja isegi langeb natuke. Optimaalseks mudeliks osutub antud juhul 7-gramm mudel.

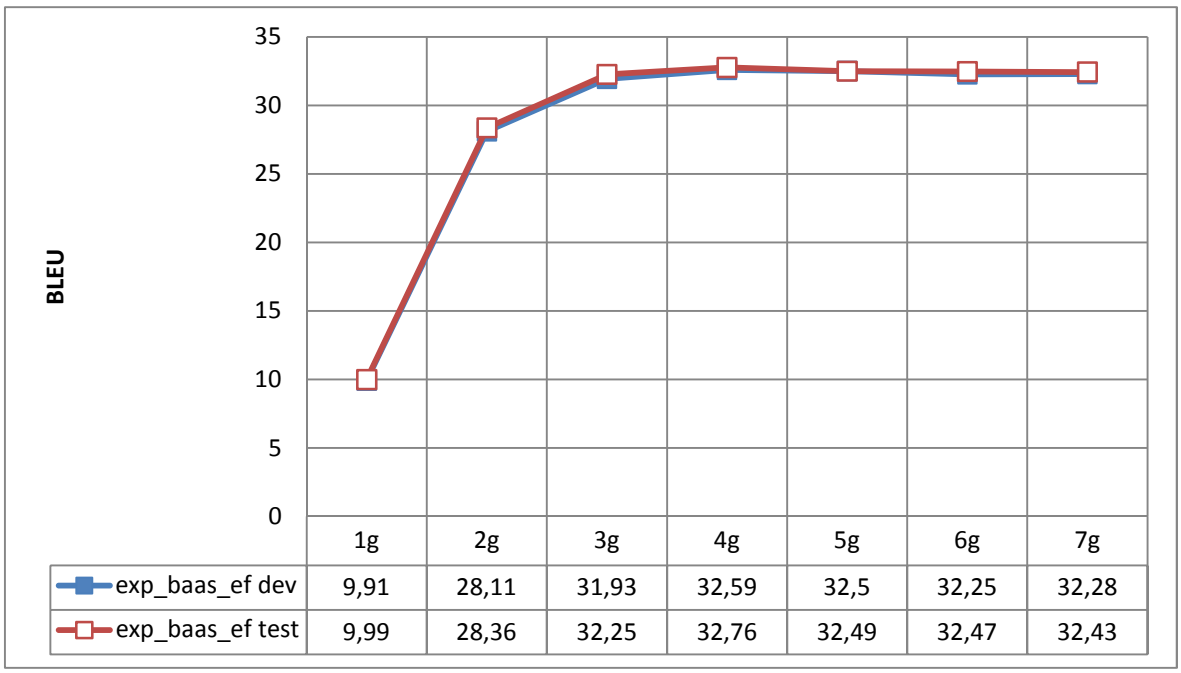


	1g	2g	3g	4g	5g	6g	7g	8g	9g
exp_baas_ee dev	19,42	35,86	42,98	45,82	46,65	46,87	46,99	46,9	46,87
exp_baas_ee test	18,41	34,18	41,69	44,27	44,81	45,18	45,04	44,99	44,93

Joonis 8. Eesti-inglise baassüsteemi tulemused 1-gramm - 9-gramm

Inglise-prantsuse korpuse korral käitus BLEU skoor keelemudeli n-gramm suuruse tõstmisel sarnaselt eelneva keelepaariga, aga optimaalne keelemudeli suurus osutus tunduvalt väiksemaks. Joonis 9 toob ära inglise-prantsuse korpusel saadud tulemused. Sealt on näha, et kui alguses tõuseb BLEU skoor tugevalt, siis umbes alates 4-gramm mudelist tõus stabiliseerub, ning isegi langeb mõne komakoha võrra.

[8] Töörühma server aadressil: liina.at.mt.ut.ee



	1g	2g	3g	4g	5g	6g	7g
exp_baas_ef dev	9,91	28,11	31,93	32,59	32,5	32,25	32,28
exp_baas_ef test	9,99	28,36	32,25	32,76	32,49	32,47	32,43

Joonis 9. Inglise-prantsuse baassüsteemi tulemused 1-gramm - 7-gramm

Katsete tulemustest lähtuvalt võtame eesti-inglise korpuse korral baassüsteemi keelemudeliks 7-gramm mudeli (BLEU skoorid: 46,99 arenduskorpusel ja 45,04 testimiskorpusel) ja inglise-prantsuse korpuse korral 4-gramm keelemudeli (BLEU skoorid: 32,59 arenduskorpusel ja 32,76 testimiskorpusel).

Lisaks selgus katsetest, et erinevate korpuste ja keelepaaride korral võib baassüsteemi optimaalne keelemudel olla väga erineva suurusega, seega tundub mõistlik see iga uue korpuse või keelepaari kasutuselevõtu korral uuesti üle katsetada, mitte kasutada keelemudeliga seotud eksperimentides eelnevates katsetes optimaalseks osutunud või rakenduste vaikimisi pakutud suurusi. See järeldus peab eriti tugevalt paika meie kasutada olnud eesti-inglise korpuse puhul, kus rakenduse vaikeväärtus (3-gramm keelemudel) oli märgatavalt kehvema tulemusega kui optimaalseks osutunud 7-gramm keelemudel. Inglise-prantsuse korpuse korral vaikeväärtus ja optimaalne väärtus nii tugevalt ei erinenud, samuti polnud väga suur nende BLEU skooride erinevus.

2.4.2 Sõnavormidest koosnev teine faktor

Põhieksperimentidest esimesena sai tehtud katse, kus baasmudelile oli lisatud teine keelemudel, mis polnud loodud sõnavormide, vaid sõnaliikide peal. Sõnaliikide genereerimiseks sihtkeele korpusesse võeti kasutusele TreeTagger märgendaja. Parema ülevaate saamiseks

sai sekundaarne keelemudel loodud erinevate suurustega, aga primaarseks keelemudeliks võeti baassüsteemis kasutatud keelemudel.

Allolev tabel 3 illustreerib antud eksperimendis kasutatud kahe faktoriga sihtkeele korpust.

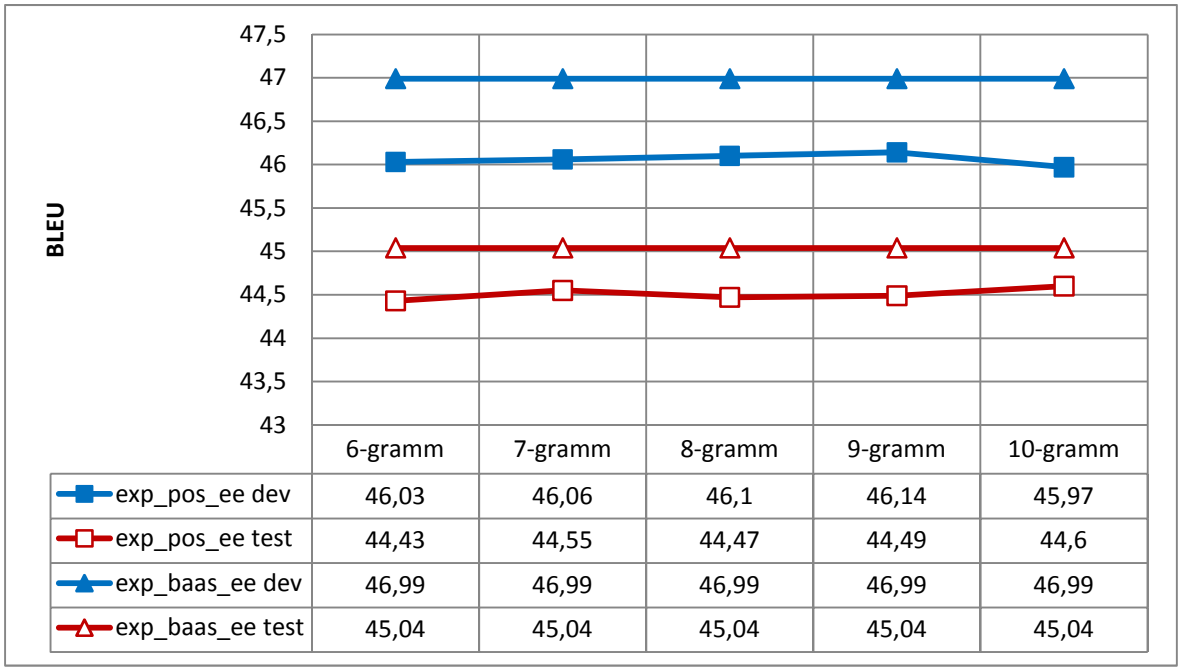
Tabel 3. Näide – kahe faktoriga korpus, sõnavormid esimeses, sõnaliigid teises faktoris

this|**DT** article|**NN** is|**VBZ** written|**VBN** by|**IN** the|**DT** european|**JJ** commission|**NN** .|**SENT**

Käesoleva eksperimendi eesmärgiks oli tekitada sihtkeelde sõnaliikidest koosnev teine faktor, ning katsetada, kas antud faktoril loodud primaarsest keelemudelist suurem keelemudel aitab kaasa baassüsteemi tõlke kvaliteedi paranemisele. Suuremaid keelemudeleid on võimalik sõnaliikide pealt luua just seetõttu, et neid on piiratud arv, seega antud faktorit kasutades on korduvaid fraase märksa rohkem kui sõnavormide korral.

Käesolevas töös on antud eksperimendi tähiseks *exp_pos_ee* (eesti-inglise korpuse puhul) või *exp_pos_ef* (inglise-prantsuse korpuse puhul).

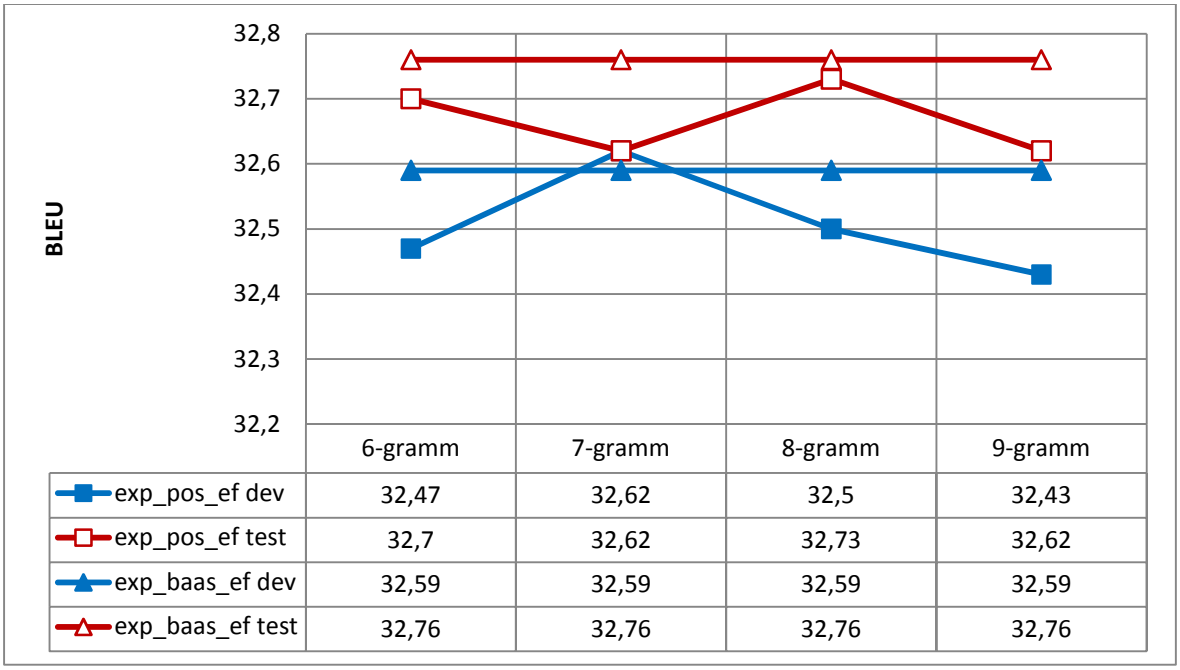
Joonis 10 toob ära eesti-inglise korpusel saadud tulemused võrrelduna baassüsteemi tulemusega. Tuleb rõhutada, et erinevalt baassüsteemist on antud eksperimendi korral primaarne keelemudel (sõnavormidel) 7-gramm suurusega, varieerub ainult sekundaarse (sõnaliikidel loodud) keelemudeli suurus.



	6-gramm	7-gramm	8-gramm	9-gramm	10-gramm
exp_pos_ee dev	46,03	46,06	46,1	46,14	45,97
exp_pos_ee test	44,43	44,55	44,47	44,49	44,6
exp_baas_ee dev	46,99	46,99	46,99	46,99	46,99
exp_baas_ee test	45,04	45,04	45,04	45,04	45,04

Joonis 10. Eesti-inglise *exp_pos* tulemused

Joonis 11 toob ära inglise-prantsuse korpusel saadud tulemused võrrelduna baassüsteemi tulemusega.



	6-gramm	7-gramm	8-gramm	9-gramm
exp_pos_ef dev	32,47	32,62	32,5	32,43
exp_pos_ef test	32,7	32,62	32,73	32,62
exp_baas_ef dev	32,59	32,59	32,59	32,59
exp_baas_ef test	32,76	32,76	32,76	32,76

Joonis 11. Inglise-prantsuse *exp_pos* tulemused

Nagu näha, on erinevate tõlkesuundade tulemused erinevad. Eesti-inglise suuna korral on uus eksperiment selgelt madalamate tulemustega kui baaseksperiment. Seega sõnaliikide peal loodud teise keelemudeli lisamine halvendas baassüsteemi tulemust. Samas vaadates inglise-prantsuse tõlkesuunda, siis on näha, et sõnaliikide peal loodud sekundaarse keelemudeli lisamine andis erinevate suuruste korral erinevaid tulemusi. Suurusega 6, 8 ja 9-gramm keelemudeli kasutamisel olid tulemused halvemad kui baassüsteemi korral, aga 7-gramm teise keelemudeli korral isegi natuke paremad kui baassüsteemi korral.

Kui eesti-inglise korral käituvad testimiskorpuse tulemused sarnaselt arenduskorpuse tulemustele, siis inglise-prantsuse korpuse korral annab testimiskorpuse kasutamine arenduskorpusest natuke erinevad tulemused, skoorid jäävad kõikide suuruste korral alla baassüsteemide skooride. Seega võib oletada, et 7-gramm keelemudeli erinevus teistest oli pigem juhuslik.

2.4.3 Sõnaklasside alusel loodud teine faktor

Teise eksperimendina sai tehtud katse, kus baasmudelile sai lisatud teine keelemudel, mis oli loodud sõnaliikidest ja sõnavormidest koosneva faktori peal. Sõnaliikide saamiseks sihtkeele korpusesse (lisafaktori loomiseks) kasutati jällegi TreeTagger märgendajat. Peale sõnaliikide sisestamist sihtkeele teise faktorisse töödeldi antud faktorit järgnevalt: kui faktori sõnaliik osutas, et antud sõna on suletud sõnade klassist, siis kopeeriti esimesest faktorist teise faktorisse antud suletud klassi sõna sõnavorm; kui sõnaliigi andmetel oli tegemist avatud klassi sõnaga, siis teist faktorit ei modifitseeritud. Sellise töötluse tulemusena oli sihtkeelses korpuses kaks faktorit, ning teises faktoris oli olenevalt sõna klassist kas sõnavorm (suletud sõnade puhul) või sõnaliik (avatud sõnade puhul). Allolev tabel 4 illustreerib kirjeldatut.

Tabel 4. Näide - sõnaklasside järgi loodud teine faktor

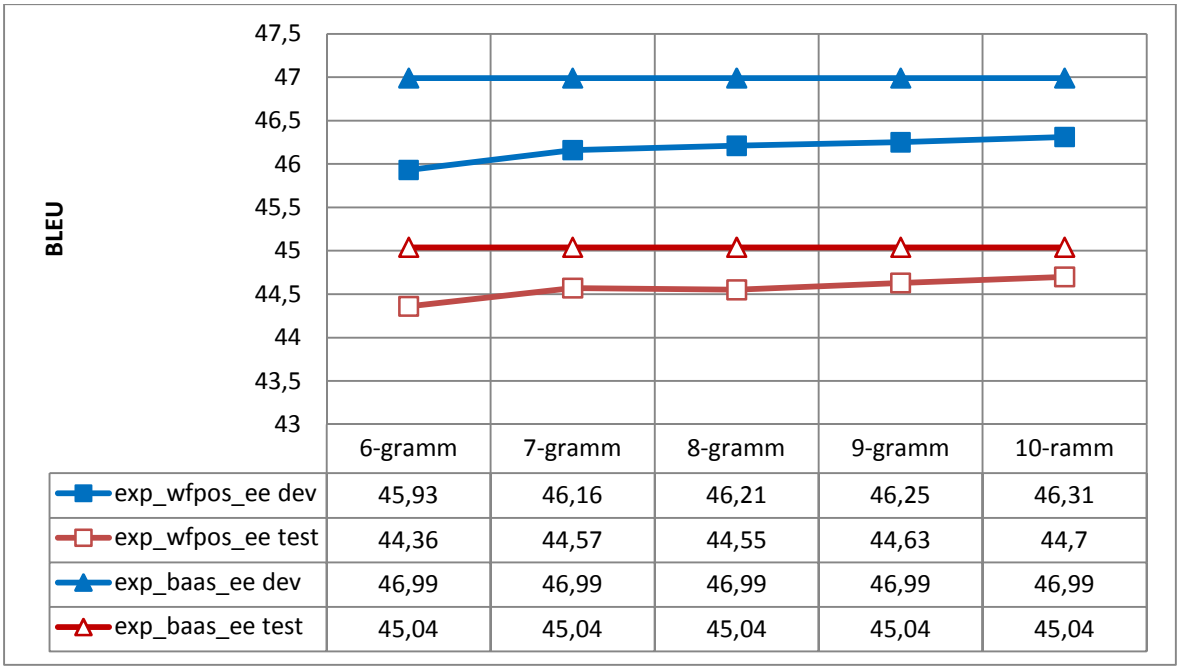
this|this article|NN is|VBZ written|VBN **by|by the|the** european|JJ commission|NN **.|.**

Käesoleva eksperimendi eesmärgiks oli tekitada sihtkeele teise faktorisse tihedalt esinevatest sõnadest ja harvem esinevate sõnade sõnaliikidest koosnevaid lausemustreid, mis ise esinedes piisavalt tihti treeningandmetes paremini modelleeriks sihtkeele lause-ehitust, ning võimaldaks luua esimesest keelemudelist suurema n-gramm suurusega keelemudeleid.

Käesolevas töös on antud eksperimendi tähiseks *exp_wfpos_ee* (eesti-inglise korpuse puhul) või *exp_wfpos_ef* (inglise-prantsuse korpuse puhul) ja teise faktori lisatöötluseks kasutati skripti *openAndClosedTaggerPos.py*[9].

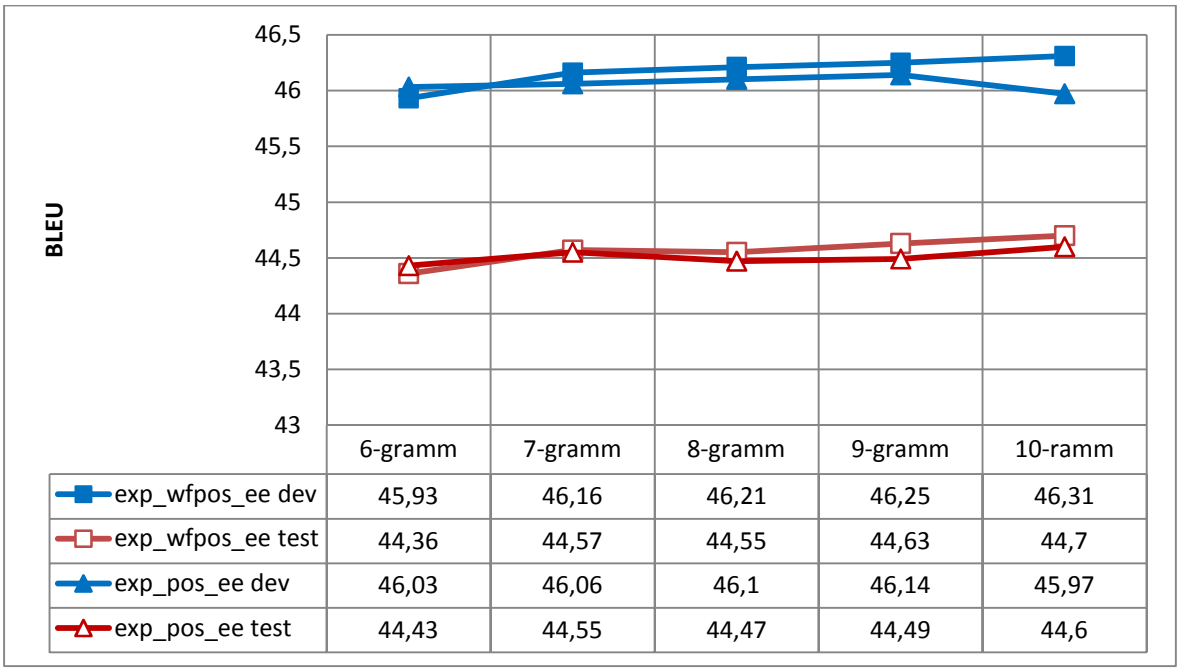
Joonis 12 toob ära eksperimendi tulemused eesti-inglise tõlkesuunal võrreldуna baas-süsteemi tulemustega. On näha, et sõnaklasside lausel teise faktori loomine annab baassüsteemist halvemad tulemused.

[9] Nimetatud skripti on lähemalt kirjeldatud töö lisas.



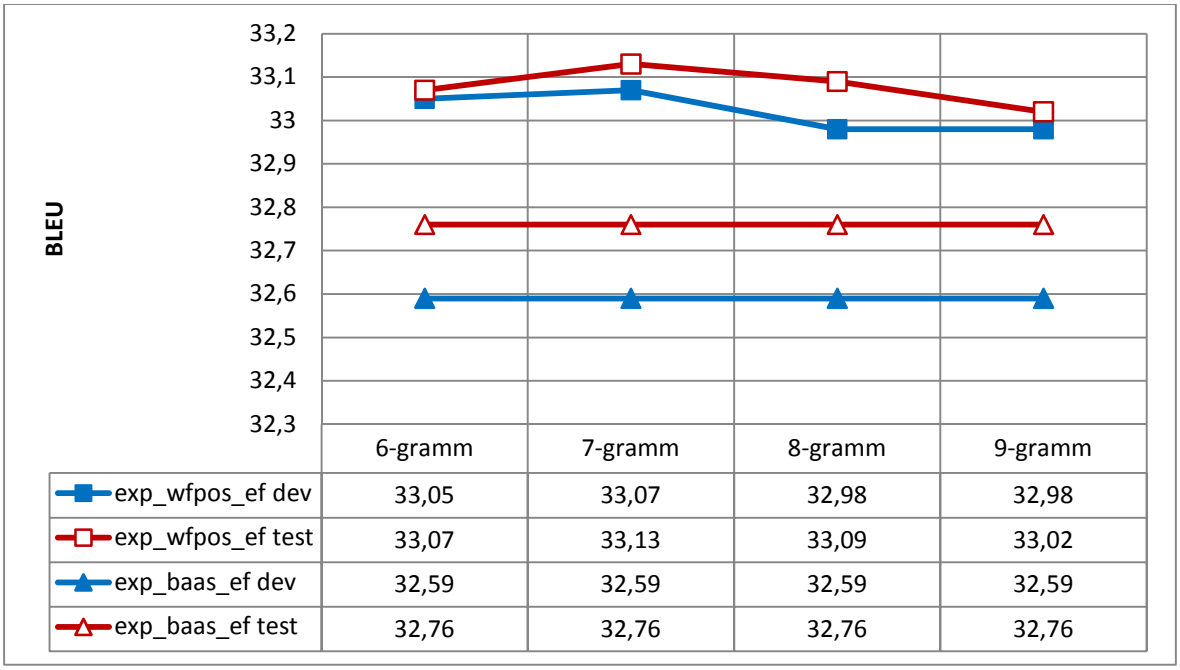
Joonis 12. Eesti-inglise *exp_wfpos* tulemused võrdluses baassüsteemi tulemustega

Joonis 13 toob ära samade tulemuste võrdluse eelneva eksperimendiga. Kuigi jooniselt lähtudes on näha, et käesolev eksperiment enamike teise keelemudeli suuruste korral andnud natuke paremaid tulemusi kui eelmine, siis on paranemine ikkagi suhteliselt väike.



Joonis 13. Eesti-inglise *exp_wfpos* tulemused võrdluses *exp_pos* tulemustega

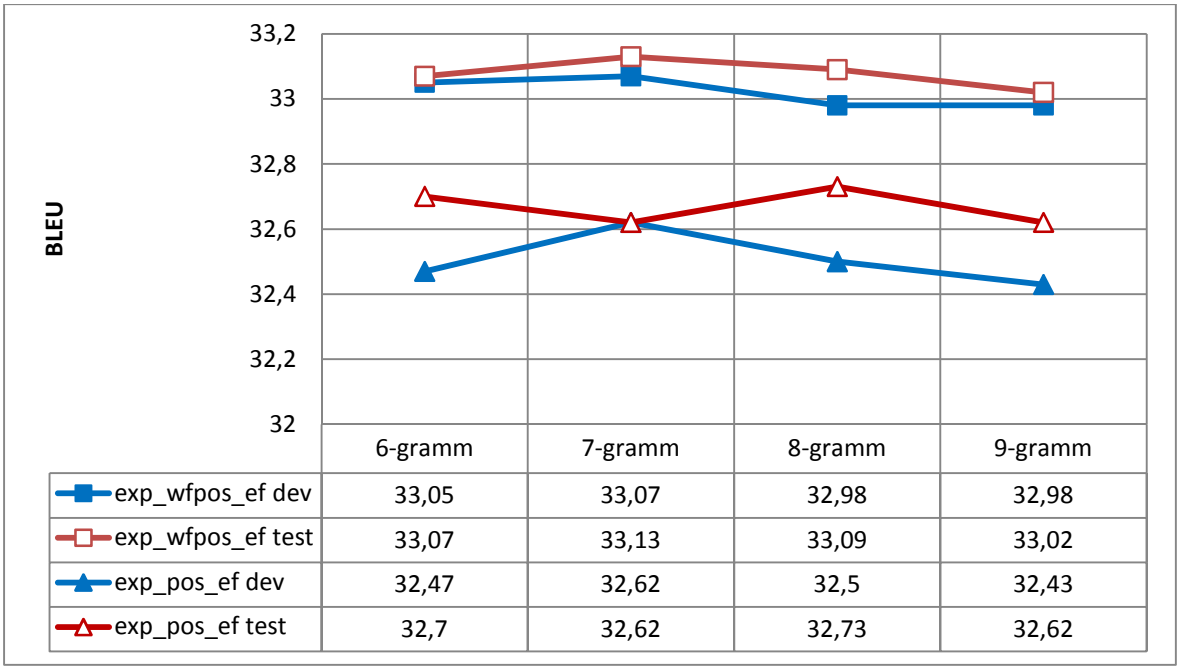
Joonis 14 sisaldab inglise-prantsuse tõlkesuuna tulemusi sama eksperimendiga võrreldes selle suuna baassüsteemiga. Erinevalt eesti-inglise tõlkesuunast, on siin näha, et käesolev lähenemine annab paremad tulemused kui baassüsteemil saadu.



	6-gramm	7-gramm	8-gramm	9-gramm
exp_wfpos_ef dev	33,05	33,07	32,98	32,98
exp_wfpos_ef test	33,07	33,13	33,09	33,02
exp_baas_ef dev	32,59	32,59	32,59	32,59
exp_baas_ef test	32,76	32,76	32,76	32,76

Joonis 14. Inglise-prantsuse *exp_wfpos* tulemused võrdluses baassüsteemi tulemustega

Allolev joonis 15 võrdleb selle eksperimendi tulemusi eelneva, *exp_pos* eksperimendiga. Ka siin on erinevalt eesti-inglise tõlkesuunast näha selgeid erinevusi, käesoleva inglise-prantsuse suuna eksperimendi tulemused on kõrgemate BLEU skooridega kui eelneva omad.



	6-gramm	7-gramm	8-gramm	9-gramm
exp_wfpos_ef dev	33,05	33,07	32,98	32,98
exp_wfpos_ef test	33,07	33,13	33,09	33,02
exp_pos_ef dev	32,47	32,62	32,5	32,43
exp_pos_ef test	32,7	32,62	32,73	32,62

Joonis 15. Inglise-prantsuse *exp_wfpos* tulemused võrdluses *exp_pos* tulemustega

Eksperimendi *exp_wfpos* kokkuvõtteks võib öelda, et erinevatel tõlkesuundadel saadi erinevad tulemused. Kui eest-inglise tõlkesuunal ei andnud see eksperiment suurt parandust ei võrreldes eelneva (*exp_pos*) eksperimendiga ning samuti jäi alla baassüsteemi skoorile, siis inglise-prantsuse tõlkesuuna korral oli tulemus vastupidine, saadud BLEU punktid olid pigem paremad nii baassüsteemi kui ka eelneva (*exp_pos*) eksperimendi omadest.

2.4.4 Sõnaklasside ja esinemissageduste alusel loodud teine faktor

Kolmandas põhieksperimendis sai edasi arendatud teise katse (*exp_wfpos*) põhimõtet. Aga kui teises eksperimendis jagati teine faktor sõnavormideks ja –liikideks sõnaklasside järgi, eeldades, et suletud klassi sõnad esinevad korpuses sagedalt, ja suletud klassi omad mitte, siis antud eksperiment püüab natuke rohkem arvestada olemasolevate korpuste eripärasid. Seega ei jagata sõnu hulkadesse mitte enam ainult sõnaklasside alusel, vaid vaadatakse ka sõnade esinemissagedusi treeningkorpuses. Teist faktorit ei muudeta sõnavormiks mitte ainult suletud klassi sõnadel, vaid ka lisaks ka *1* .. *n* sagedasemalt esineva sõna puhul.
Illustreerimiseks oletame, et tegeleme eelmise eksperimendi juurest toodud lausenäitega (tabel 4), ja suuruseks *n* valime näiteks 50. Seega, kui luua eesti-inglise tõlkesuuna inglis-keelsest korpuseosast kahanevalt sorteeritud unigrammide esinemissagedusi sisaldav

nimekiri[10], siis näeme, et avatud klassi sõnad „*article*“, „*european*“ ning „*commission*“ on viiekümne tihedamini esineva sõna seas, ning vastava lause teine faktor näeks välja nagu kujutab allolev tabel 5.

Tabel 5. Näide - sõnaklasside ja sõnade esinemissageduste järgi loodud teine faktor.

this|this article|article is|VBZ written|VBN **by|by the|the european|european commission|commission .|.**

Antud eksperimendi eesmärgiks on sarnaselt eelmise eksperimendiga luua sihtkeele korpusesse selline teine faktor, mis moodustaks sagedasti esinevaid sõnemustreid, ja võimaldaks saada sekundaarse keelemudeli loomiseks rohkem näiteid.

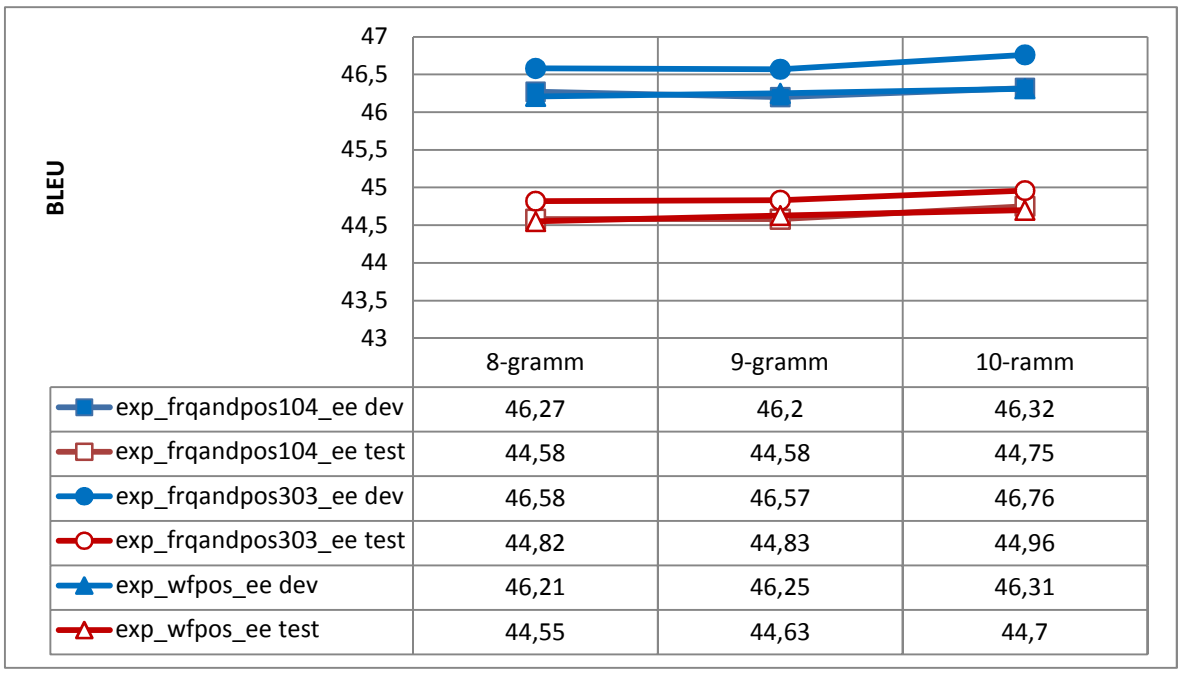
Käesolevas töös on antud eksperimendi tähiseks „*exp_frqandpos[n]_ee*“ (eesti-inglise korpuse puhul) või „*exp_frqandpos[n]_ef*“ (inglise-prantsuse korpuse puhul) ja teise faktori lisatöötluseks kasutati skripte *openAndClosedTaggerPos.py* ja *openAndClosedTaggerFrq.py*[11].

Märgendi osa *n* tähistab numbrit, mis näitab, mitu sagedusnimekirja esimest sõna loeti suletud sõnadega samasse hulka. Kuna selle suuruse valik toimus sagedusnimekirjade heuristilisel uurimisel, siis sai valitud mõlema tõlkesuuna jaoks kaks erinevat suurust, püüdes valikud teha vastavalt suletud klassi sõnade esinemissagedustele sagedusnimekirjas. Eesti-inglise tõlkesuuna korral sai valitud suurused 104 sõna ja 303 sõna sagedusnimekirjast, inglise-prantsuse suuna korral suurused 218 ja 403 sõna sagedusnimekirjast.

Allolev joonis 16 toob ära antud eksperimendi eesti-inglise tõlkesuuna tulemused võrdluses eelmise eksperimendi tulemustega. Eksperiment *exp_frqandpos303* sai natuke paremaid tulemusi kui eksperiment *exp_frqandpos104*. Üks tõlgendus antud tulemuse kohta on see, et antud tõlkesuuna korral tõlkimisel eelistatakse, et teises keelemudelis oleks võimalikult palju sõnavorme.

Võrreldes eelneva eksperimendiga, on *exp_frqandpos104* tulemused väga sarnased *exp_wfpos* tulemustele, samas *exp_frqandpos303* tulemused on natuke paremad. Kuid ka selle eksperimendi tulemused ei ületa antud tõlkesuuna baassüsteemi tulemusi.

[10] Sagedusnimekirjade näited on ära toodud 2. lisa juures.
[11] Nimetatud skripti on lähemalt kirjeldatud töö 1. lisas.

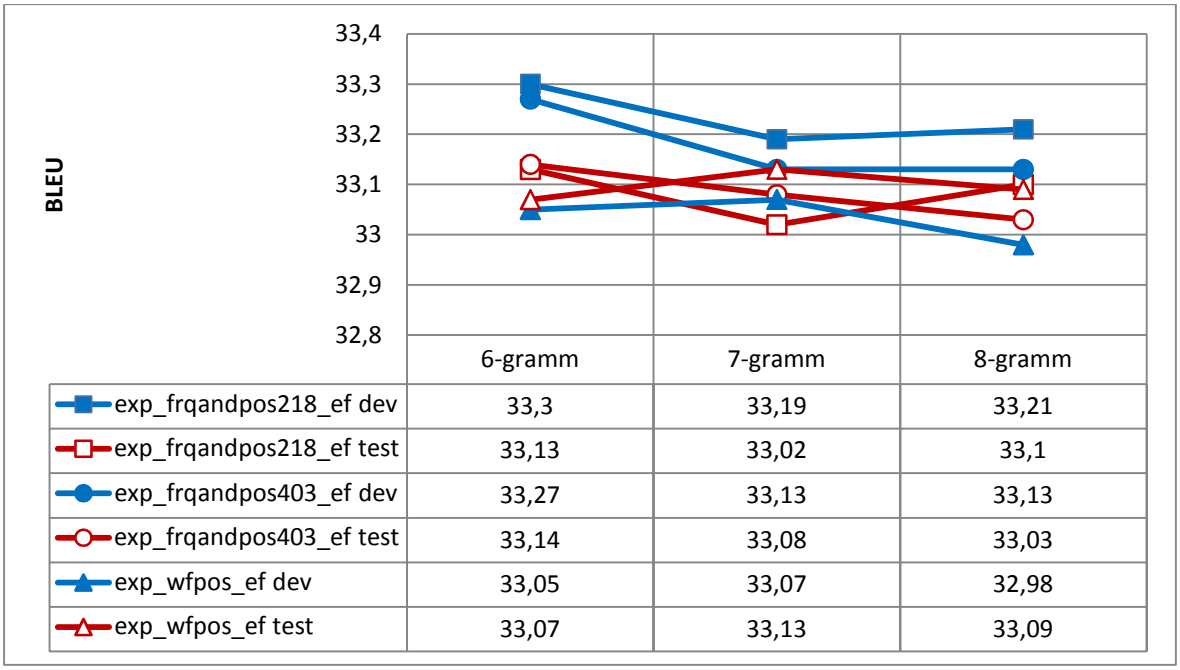


	8-gramm	9-gramm	10-ramm
exp_frqandpos104_ee dev	46,27	46,2	46,32
exp_frqandpos104_ee test	44,58	44,58	44,75
exp_frqandpos303_ee dev	46,58	46,57	46,76
exp_frqandpos303_ee test	44,82	44,83	44,96
exp_wfpos_ee dev	46,21	46,25	46,31
exp_wfpos_ee test	44,55	44,63	44,7

Joonis 16. Eesti-inglise *exp_frqandpos104* ja *exp_frqandpos303* tulemused võrdluses *exp_wfpos* tulemustega

Inglise-prantsuse suuna tulemusi antud eksperimendi kohta kujutab allolev joonis 17. Nagu jooniselt näha, andis selle eksperimendi strateegia paremad tulemused kui eelneva, *exp_wfpos* oma. Seega on tulemused ka jätkuvalt paremad kui antud suuna baassüsteemi tulemused.

Võrreldes eksperimente *exp_frqandpos218* ja *exp_frqandpos403* omavahel, on näha, et tulemused on üldiselt sarnased, ainult 8-gramm suurusega keelemudeli korral on esimene saanud viimatimainitust märgatavalt paremaid tulemusi.



	6-gramm	7-gramm	8-gramm
exp_frqandpos218_ef dev	33,3	33,19	33,21
exp_frqandpos218_ef test	33,13	33,02	33,1
exp_frqandpos403_ef dev	33,27	33,13	33,13
exp_frqandpos403_ef test	33,14	33,08	33,03
exp_wfpos_ef dev	33,05	33,07	32,98
exp_wfpos_ef test	33,07	33,13	33,09

Joonis 17. Inglise-prantsuse *exp_frqandpos218* ja *exp_frqandpos403* tulemused võrdluses *exp_wfpos* tulemustega

Sõnaliikidest ja sõnavormidest koosneva ning sõnaklasside ja esinemissageduste alusel loodud teise faktori eksperimendi tulemuste kokkuvõtteks võib öelda, et eesti-inglise tõlkesuuna puhul olid küll antud eksperimendi tulemused natuke paremad kui eelneva omad, aga samas jäid ikka alla selle tõlkesuuna baassüsteemi tulemustele.

Inglise-prantsuse suuna puhul olid skoorides väikesed paranemised võrreldes eelneva eksperimendiga, ning skoorid ületasid jätkuvalt ka baassüsteemi skoori.

2.4.5 Ainult esinemissageduste alusel loodud teine faktor

Neljandas eksperimendis sai läbi viidud veel üks variatsiooni teisest (*exp_wfpos*) ja kolmandast (*exp_frqandpos*) eksperimendist. Kui teises eksperimendis jagati teine faktor sõnavormideks ja –liikideks sõnaklasside järgi, ja kolmas eksperiment püüdis natuke rohkem arvestada olemasolevate korpuste eripärasid, käitudes sagedaste sõnade teise faktoriga sarnaselt suletud sõnadega, siis seekord jagame sõnad kahte hulka ainult nende esinemissageduste järgi. Seega teist faktorit ei muudeta sõnavormiks mitte suletud klassi sõnadel, vaid *1* .. *n* sagedasemalt esineva sõna puhul.

Paremaks illustreerimiseks oletame, et tegeleme eelmist eksperimentide juures toodud lausenäidetega (tabel 4 ja tabel 5). Kui *n* (antud näite puhul *n=50*) sagedasema sõna seas on sõnad ja märgendid:

- „*by*“,
- „*european*“,
- „*commission*“,
- „.“,

siis näitelause teine faktor kujuneb nagu illustreerib tabel 6:

Tabel 6. Näide - sõnade esinemissageduste järgi loodud teine faktor.

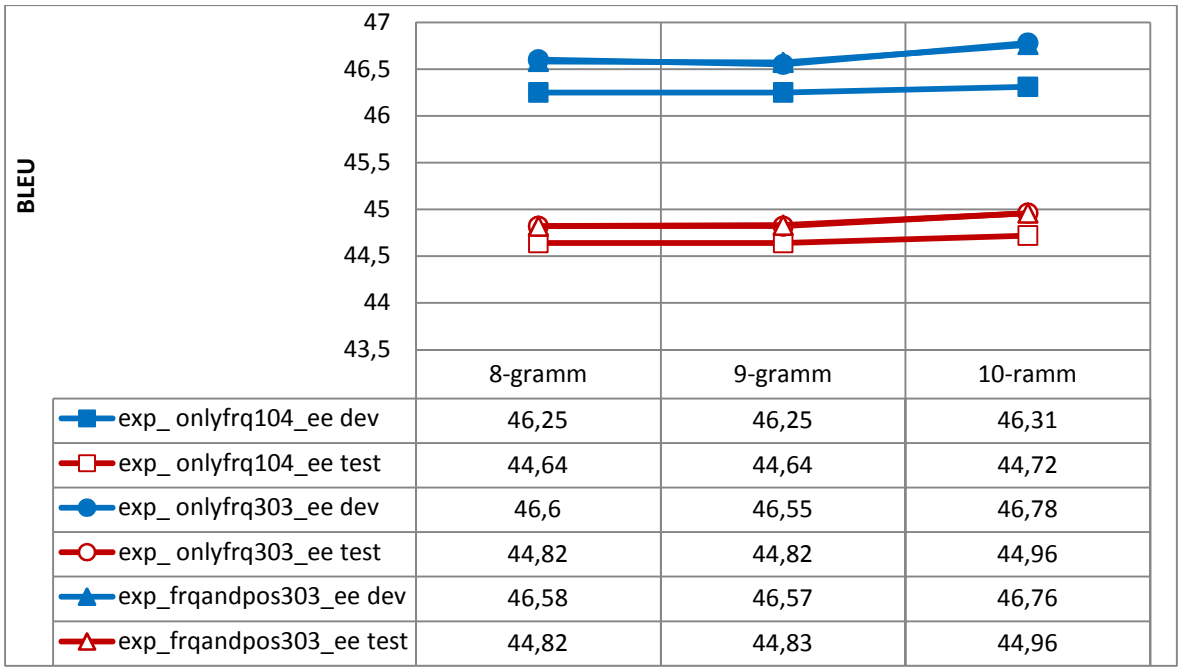
this|DT article|NN is|VBZ written|VBN **by|by** the|DT **european|european commission|commission .|.**

Antud eksperimendi eesmärgiks on sarnaselt eelmistega luua sihtkeele korpusesse selline teine faktor, mis moodustaks sagedasti esinevaid mustreid ja võimaldaks saada sekundaarse keelemudeli loomiseks rohkem ühesuguseid näiteid.

Käesolevas töös on antud eksperimendi tähiseks „*exp_ onlyfrq[n]_ee*“ (eesti-inglise korpuse puhul) või „*exp_onlyfrq[n]_ef*“ (inglise-prantsuse korpuse puhul) ja teise faktori lisatöötluseks kasutati skripti *openAndClosedTaggerFrq.py*.

Märgendi osa *n* tähistab numbrit, mis näitab, mitu sagedusnimekirja esimest sõna loeti esimesse hulka. Need suurused said valitud samad, mis eelneva (exp_frqandpos) eksperimentide juures: eesti-inglise tõlkesuunal 104 ja 304, inglise-prantsuse tõlkesuunal 218 ja 403.

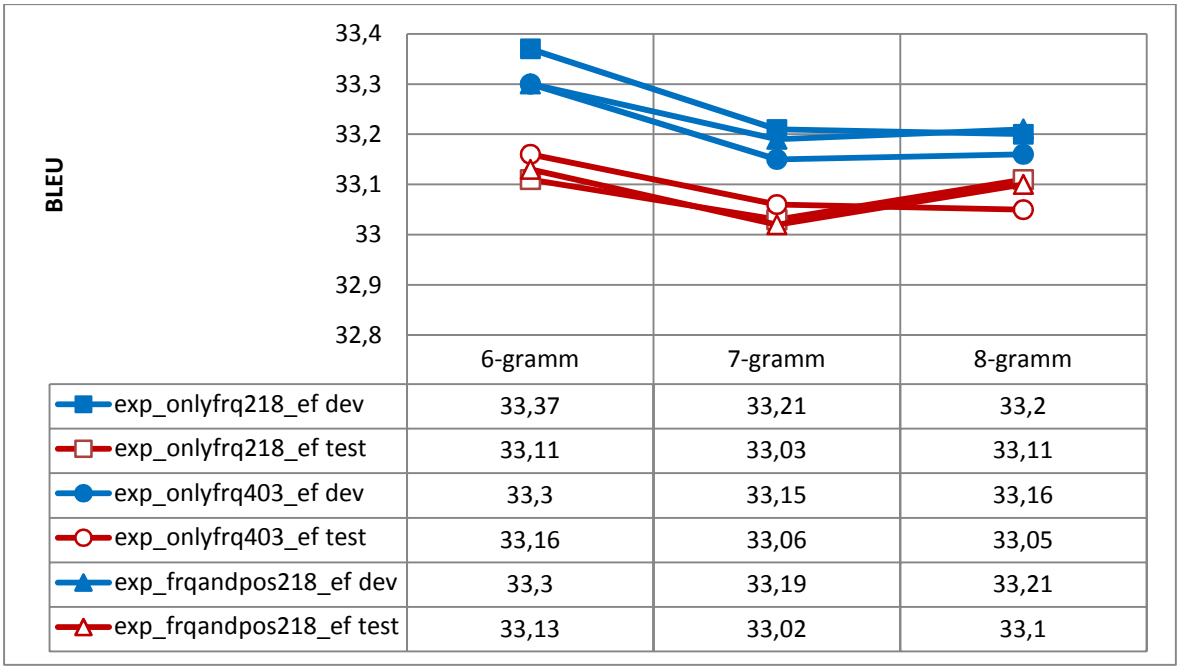
Võrreldes eesti-inglise tõlkesuuna eksperimendi tulemusi eelneva eksperimendi parimate tulemustega (*exp_frqandpos303*), on allolevalt graafikult (joonis 18) näha, et *exp_frqandpos303* ja *exp_onlyfrq303* on saanud peaaegu identsed tulemused. Samuti saab samuti öelda, et ka selle eksperimendi tulemused ei ületanud antud tõlkesuuna baassüsteemi tulemusi.



	8-gramm	9-gramm	10-ramm
exp_ onlyfrq104_ee dev	46,25	46,25	46,31
exp_ onlyfrq104_ee test	44,64	44,64	44,72
exp_ onlyfrq303_ee dev	46,6	46,55	46,78
exp_ onlyfrq303_ee test	44,82	44,82	44,96
exp_frqandpos303_ee dev	46,58	46,57	46,76
exp_frqandpos303_ee test	44,82	44,83	44,96

Joonis 18. Eesti-inglise *exp_onlyfrq104* ja *exp_onlyfrq303* tulemused võrdluses *frqandpos303* tulemustega

Vaadeldes inglise-prantsuse eksperimendi tulemusi võrdluses sama suuna eelmise eksperimendiga (joonis 19), on näha, et see eksperiment (erinevalt eelmistest antud suuna omadest) ei andnud võrreldes eelnevalt tehtud eksperimendiga eriti paremaid tulemusi. Samas on tulemused jätkuvalt paremad kui baassüsteemi omad.



	6-gramm	7-gramm	8-gramm
exp_onlyfrq218_ef dev	33,37	33,21	33,2
exp_onlyfrq218_ef test	33,11	33,03	33,11
exp_onlyfrq403_ef dev	33,3	33,15	33,16
exp_onlyfrq403_ef test	33,16	33,06	33,05
exp_frqandpos218_ef dev	33,3	33,19	33,21
exp_frqandpos218_ef test	33,13	33,02	33,1

Joonis 19. Inglise-prantsuse *exp_onlyfrq218* ja *exp_onlyfrq403* tulemused võrdluses *frqandpos218* tulemustega

Eksperimendi *exp_onlyfrq* on kokkuvõtteks võib öelda, et sarnaselt eelmistele eksperimentidele andis siin kirjeldatud keelemudeli loomise tehnika kasutamine positiivseid tulemusi inglise-prantsuse tõlkesuuna korral, ning „halvemaid kui baassüsteem“ tulemusi eesti-inglise tõlkesuuna korral.

2.4.6 Juhuslikkuse alusel loodud teine faktor

Viies eksperiment sai loodud kontrollimaks eelnevate eksperimentide tulemuste olulisust. Antud eksperiment ei oma mingit kindlat motiveeritud metoodikat sõnade kahte hulka jagamiseks ning nende hulkade alusel sihtkeele korpusesse teise faktori loomiseks, vaid kasutab selleks juhuslikkust. Samas ikka jälgides, et saadud hulkadest esimeses (selles, mis oli algselt suletud klassi sõnade hulk) oleks sõnade summaarne esinemissagedus treeningkorpuse korral sama, mis eelneval, *exp_onlyfrq* eksperimendil. Seega, kui näiteks *exp_onlyfrq104_ee* korral oli nende sõnade, mille puhul pandi teise faktorisse sõnavorm, kogusageduseks *x*, siis *exp_rnd104_ee* korral on see samuti *x*.

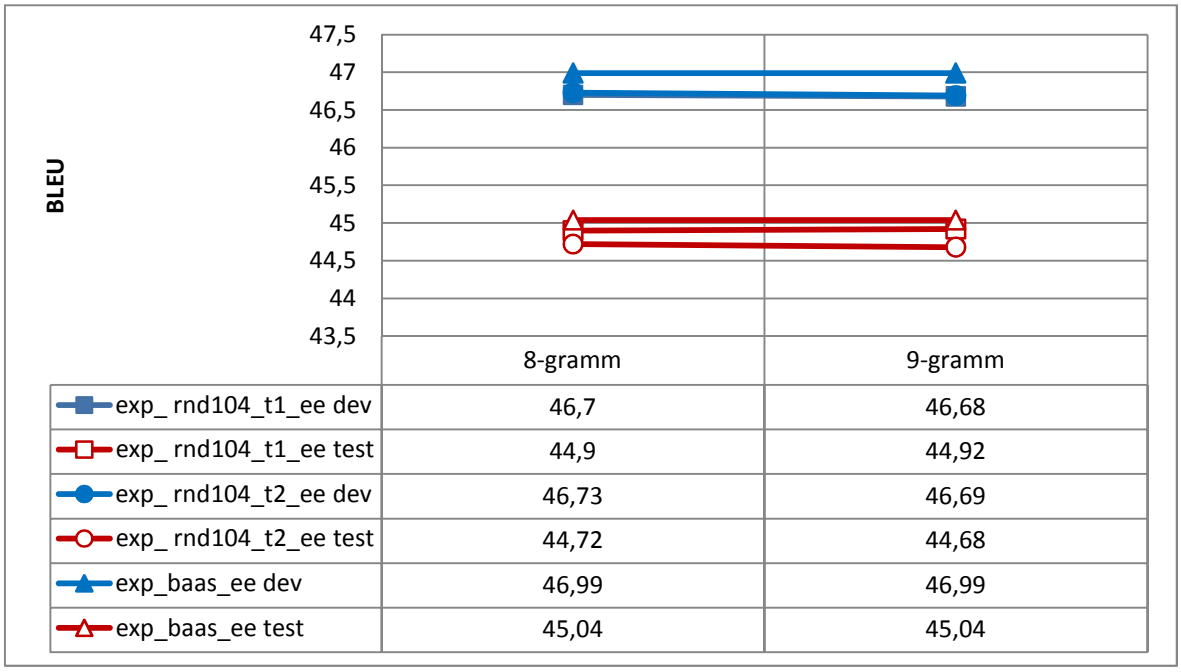
Et saada adekvaatsemaid tulemusi, teeme iga *exp_onlyfrq* eksperimendi kohta kaks (*t1* ja *t2*) *exp_rnd* eksperimenti.

Käesolevas töös on antud eksperimendi tähiseks „*exp_ rnd[n]_t[jrk]_ee*“ (eesti-inglise korpuse puhul) või „*exp_ rnd[n]_t[jrk]_ef*“ (inglise-prantsuse korpuse puhul). Juhusliku

valiku läbiviimiseks kasutati skripti *generateRndWordList.py*[12], teise faktori lisatöötluseks skripti *openAndClosedTaggerFrq.py*.

Eksperimendi põhiliseks ideeks on saada teada, kas eelnevates eksperimentides kasutatud strateegiad sekundaarse keelemudeli loomiseks olid õigustatud, või saab sama häid või isegi paremaid tulemusi lihtsalt juhuslikkuse alusel sõnu kahte hulka jagades.

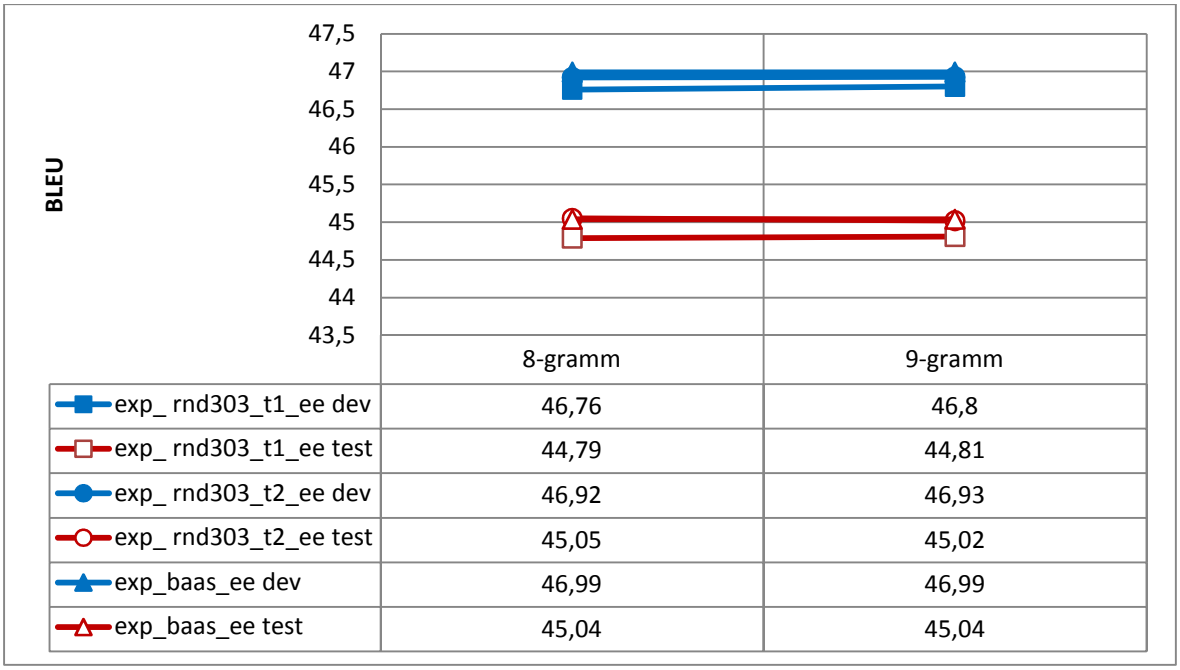
Joonis 20 ja joonis 21 illustreerivad, et eesti-inglise tõlkesuuna puhul kõik neli juhuslikkuse alusel loodud eksperimenti ei ületa oma tulemustes baassüsteemil saadud tulemusi.



	8-gramm	9-gramm
exp_ rnd104_t1_ee dev	46,7	46,68
exp_ rnd104_t1_ee test	44,9	44,92
exp_ rnd104_t2_ee dev	46,73	46,69
exp_ rnd104_t2_ee test	44,72	44,68
exp_baas_ee dev	46,99	46,99
exp_baas_ee test	45,04	45,04

Joonis 20. Eesti-inglise *exp_rnd104_t1* ja *exp_rnd104_t2* tulemused võrdluses baassüsteemi tulemustega

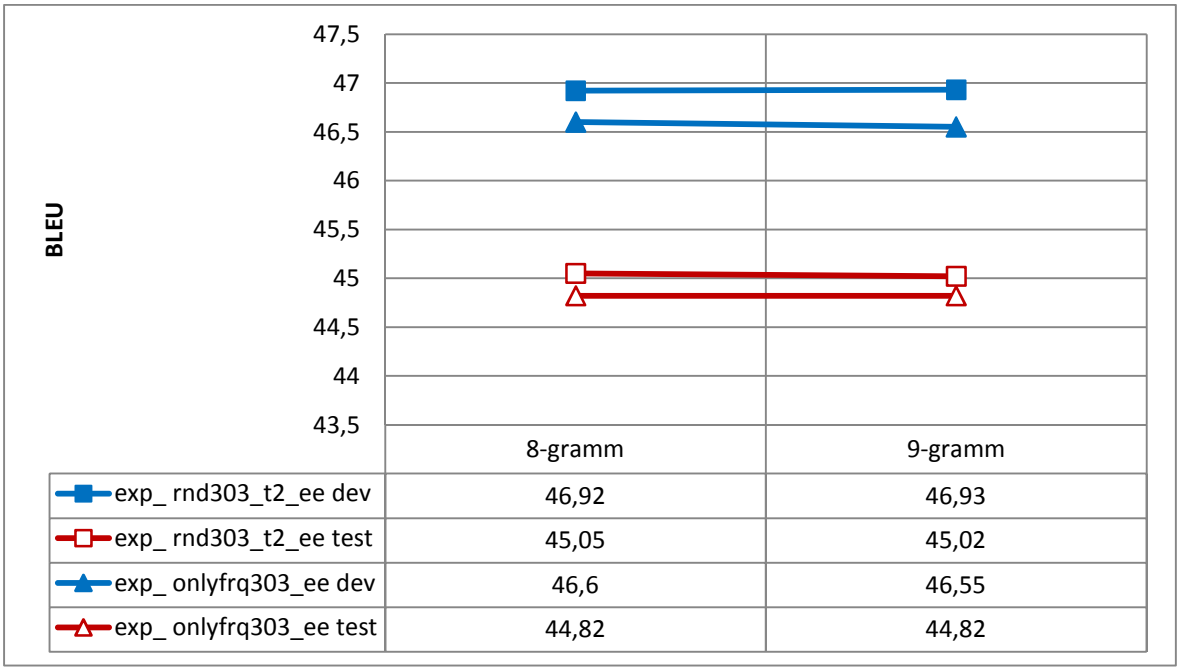
[12] Nimetatud skripti on lähemalt kirjeldatud töö 1. lisas.



	8-gramm	9-gramm
exp_ rnd303_t1_ee dev	46,76	46,8
exp_ rnd303_t1_ee test	44,79	44,81
exp_ rnd303_t2_ee dev	46,92	46,93
exp_ rnd303_t2_ee test	45,05	45,02
exp_baas_ee dev	46,99	46,99
exp_baas_ee test	45,04	45,04

Joonis 21. Eesti-inglise *exp_rnd303_t1* ja *exp_rnd303_t2* tulemused võrdluses baassüsteemi tulemustega

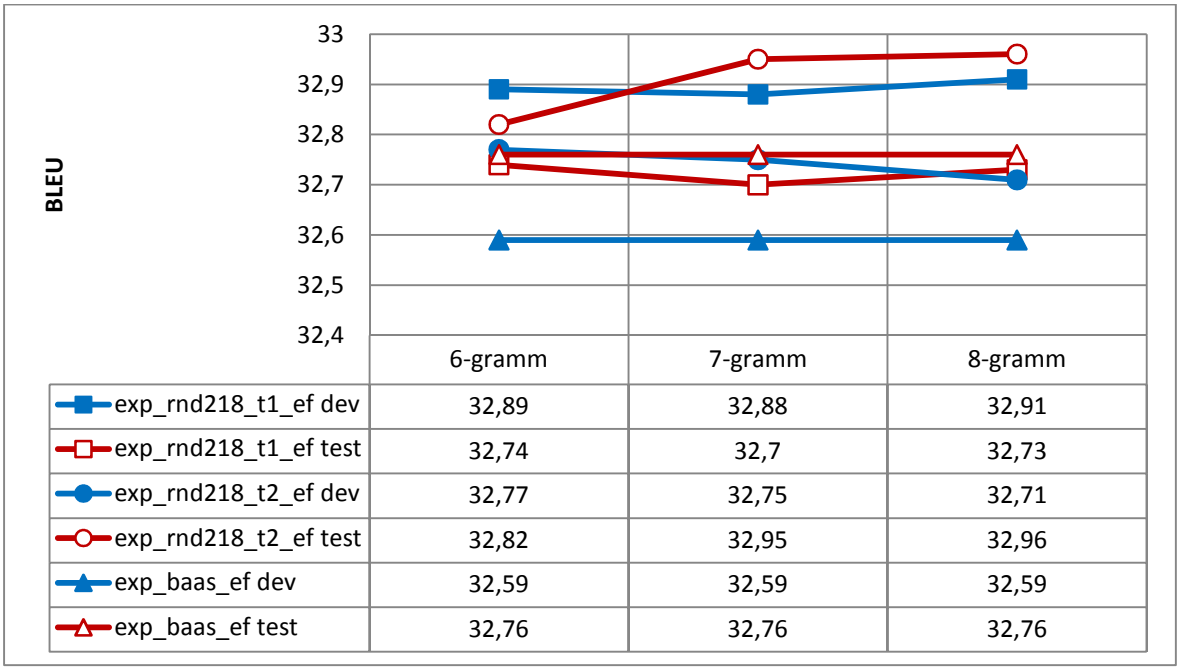
Samas, vaadates eelneva eksperimendi parimaid tulemusi ja juhuslikkuse alusel genereeritud eksperimentidest parimaid tulemusi andnud eksperimendi *exp_rnd303_t2* tulemusi (joonis 22), näeme, et eesti-inglise korpuse puhul saab juhuslikkuse alusel genereeritud eksperimendi korral isegi paremaid tulemusi kui eelnevate eksperimentide strateegiate puhul.



	8-gramm	9-gramm
exp_ rnd303_t2_ee dev	46,92	46,93
exp_ rnd303_t2_ee test	45,05	45,02
exp_ onlyfrq303_ee dev	46,6	46,55
exp_ onlyfrq303_ee test	44,82	44,82

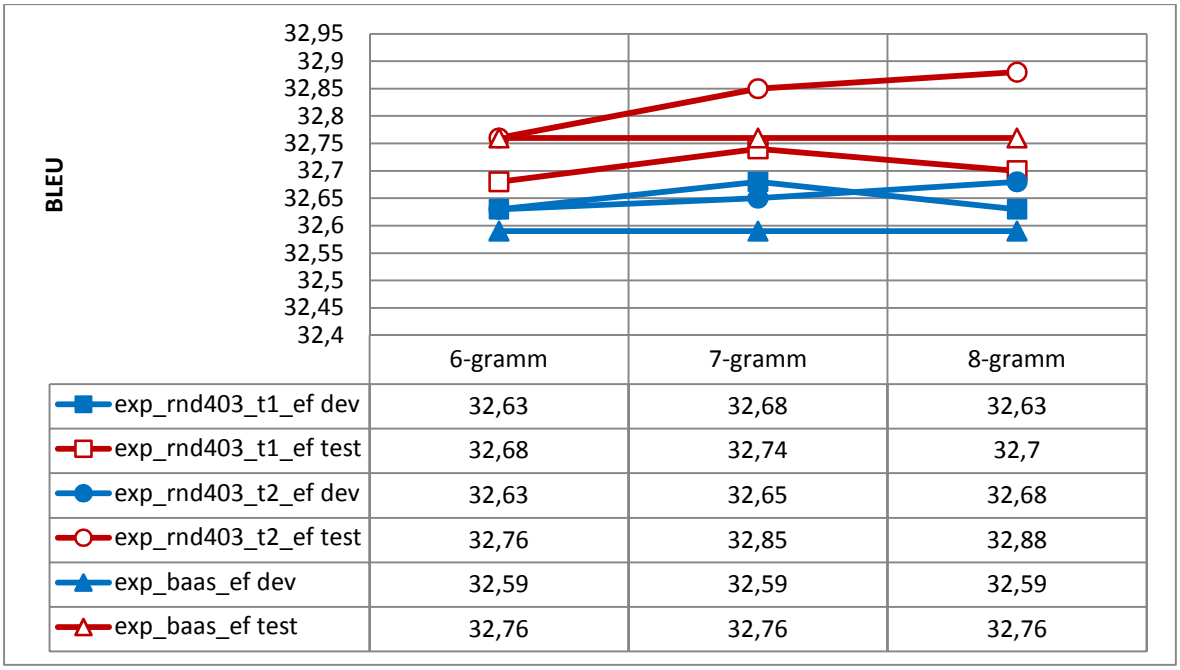
Joonis 22. Eesti-inglise *exp_rnd303_t2* ja *exp_onlyfrq303* tulemuste võrdlus

Inglise-prantsuse tõlkesuuna korral vaadeldes juhuslikkuse alusel tehtud eksperimente, ning võrreldes neid baassüsteemi tulemustega (joonis 23 ja joonis 24), näeme, et juhuslikkuse alusel loodu eksperimendid saavad ainult natuke parema või peaaegu sama skoori kui baassüsteemi eksperimendid.



	6-gramm	7-gramm	8-gramm
exp_rnd218_t1_ef dev	32,89	32,88	32,91
exp_rnd218_t1_ef test	32,74	32,7	32,73
exp_rnd218_t2_ef dev	32,77	32,75	32,71
exp_rnd218_t2_ef test	32,82	32,95	32,96
exp_baas_ef dev	32,59	32,59	32,59
exp_baas_ef test	32,76	32,76	32,76

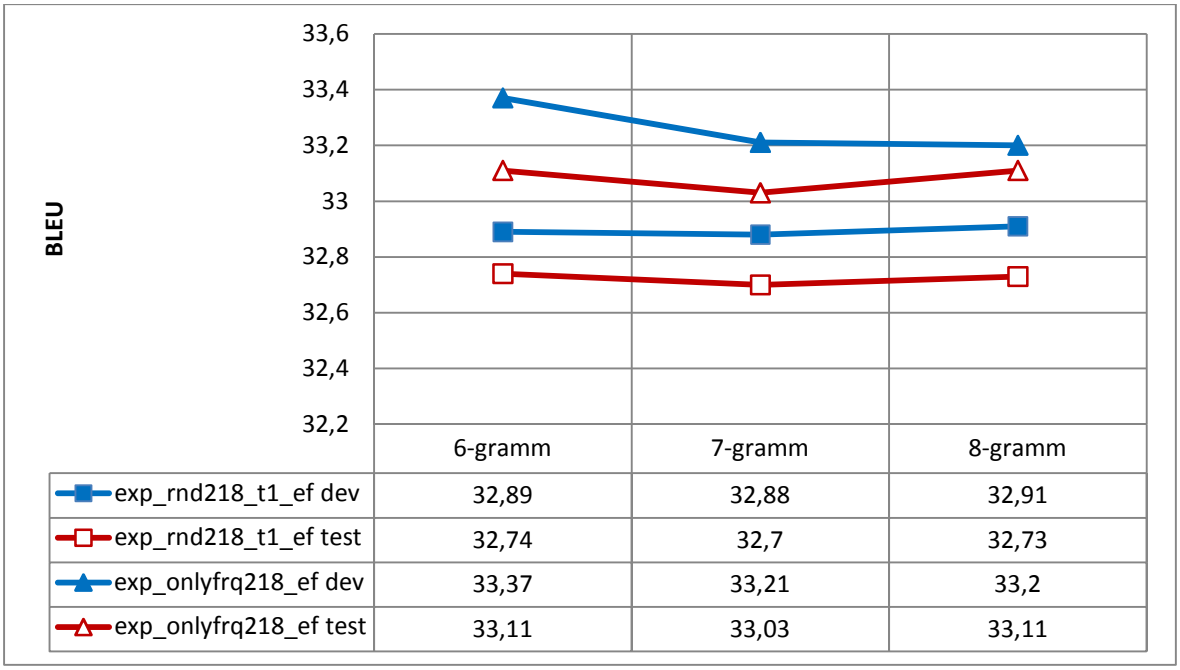
Joonis 23. Inglise-prantsuse *exp_rnd218_t1* ja *exp_rnd218_t2* tulemused võrdluses baassüsteemi tulemustega



	6-gramm	7-gramm	8-gramm
exp_rnd403_t1_ef dev	32,63	32,68	32,63
exp_rnd403_t1_ef test	32,68	32,74	32,7
exp_rnd403_t2_ef dev	32,63	32,65	32,68
exp_rnd403_t2_ef test	32,76	32,85	32,88
exp_baas_ef dev	32,59	32,59	32,59
exp_baas_ef test	32,76	32,76	32,76

Joonis 24. Inglise-prantsuse *exp_rnd403_t1* ja *exp_rnd403_t2* tulemused võrdluses baassüsteemi tulemustega

Samas, võrreldes juhuslikkuse alusel loodud mudelitest parimat mudelit (*exp_rnd218_t1*) eelnevas eksperimendis loodud parima mudeliga (*exp_onlyfrq218*), on näha (joonis 25), et antud tõlkesuuna korral annab juhuslik mudel halvema tulemuse. Seega võib öelda, et antud keelepaari korral olid eelnevates eksperimentides katsetatud strateegiad põhjendatud ning pole võimalik triviaalselt saada paremat tulemust luues sihtkeelde teise faktorit juhuslikkuse alusel.



	6-gramm	7-gramm	8-gramm
exp_rnd218_t1_ef dev	32,89	32,88	32,91
exp_rnd218_t1_ef test	32,74	32,7	32,73
exp_onlyfrq218_ef dev	33,37	33,21	33,2
exp_onlyfrq218_ef test	33,11	33,03	33,11

Joonis 25: Inglise-prantsuse *exp_rnd218_t1* ja *exp_onlyfrq218* tulemuste võrdlus

Antud eksperimendi *exp_rnd* kokkuvõtteks võib öelda, et vastavalt tulemustele annavad inglise-prantsuse tõlkesuuna korral eelnevates eksperimentides katsetatud strateegiad parema tulemuse kui baassüsteemi või juhusliku valiku kasutamisel. Samas eesti-inglise tõlkesuuna korral ei ole katsetatud strateegiad mitte ainult halvemad baassüsteemi tulemustest, vaid saadud tulemused on ületatavad ka juhusliku valiku kasutamisel. Eelnevat võib seletada asjaolu, et juhuslikkuse alusel teise faktori loomisel võetakse nn esimesse hulka (need sõnad, mille puhul teises faktoris on sõnavorm) tunduvalt suurem hulk sõnu kui sageduse põhjal loodud faktori puhul (ehk kompenseeritakse väga paljude harva esinevate sõnade kaasamisega mõne väga sagedase sõna väljajätmist). Samas on ka palju selliseid (väga sagedasi) sõnu, mis mõlemal korral satuvad esimesse hulka. Seega on juhusliku valiku korral tulemus, kus teine faktor sisaldab väga palju sõnavorme (ja suhteliselt vähe sõnaliike), ning tundub, et sellist tulemust tundub eelistavat tõlkesüsteem eesti-inglise tõlkesuuna puhul.

2.5 Tulemuste kontrollimine kasutades kolmandat tõlkesuunda

Eelnevalt sai ära toodud erinevad strateegiad teise keelemudeli loomiseks kasutades siht-keele korpusesse sisestatud lisainfot sõnaliikide näol. Tehtud eksperimentidest lähtus, et need strateegiad andsid erinevate tõlkesuundade puhul erinevaid tulemusi. Inglise-

prantsuse suuna puhul oli tõlkekvaliteedi paranemine nähtav (BLEU skoorid tõusid), samas eesti-inglise tõlkesuuna korral ei saadud paranemist.

Kirjeldatud erinevaid tulemusi erinevatel tõlkesuundadel seletab suure tõenäosusega üks järgnevast kahest võimalusest. Esiteks on võimalik, et erinevused tulemustes olid tingitud sihtkeelest ja/või kasutatud korpuse valdkonnast. Seega baassüsteemi mitte ületanud tulemused tingis kas siis inglise keel kui sihtkeel, või kasutatud JRC-Acquis korpus või mõlemad faktorid koos. Teine võimalus on, et erinevad tulemused kahe tõlkesuuna vahel olid tingitud mingitest teistest faktoritest, näiteks tõlkesuuna lähtekeelest või kasutatud korpuste suurustest vms. Selleks, et kontrollida, kumma võimalusega on tegemist, viiakse läbi lisakatseid.

Lisakatsete läbiviimiseks võetakse kasutusele kolmas tõlkesuund (edaspidi nimetatudki kui „kolmas“ tõlkesuund), tõlkides prantsuse-inglise suunal, ning treenimisandmetena kasutades JRC-Acquis korpust. Sellisel juhul kasutavad esimene ja kolmas tõlkesuund sama korpust ning sihtkeelt. Ja kui kolmanda tõlkesuuna tulemused sarnanevad esimese tõlkesuuna tulemustele, siis tulenes erinevus esimese ja teise tõlkesuuna vahel kas sihtkeelest või kasutatud korpuse valdkonnast või nende kahe faktori koosmõjust. Kui kolmanda tõlkesuuna tulemused ei sarnane esimese tõlkesuuna tulemustele, siis tõenäoliselt tulenesid erinevused esimese ja teise tõlkesuuna vahel mingitest teistest faktoritest (nagu eespool mainitud).

Korpuse ettevalmistamine kolmanda tõlkesuuna jaoks toimub sarnaselt kahele esimesele tõlkesuunale, tabel 7 toob ära prantsuse-inglise tõlkesuuna JRC-Acquis korpuse suuruse peale eeltöötlust.

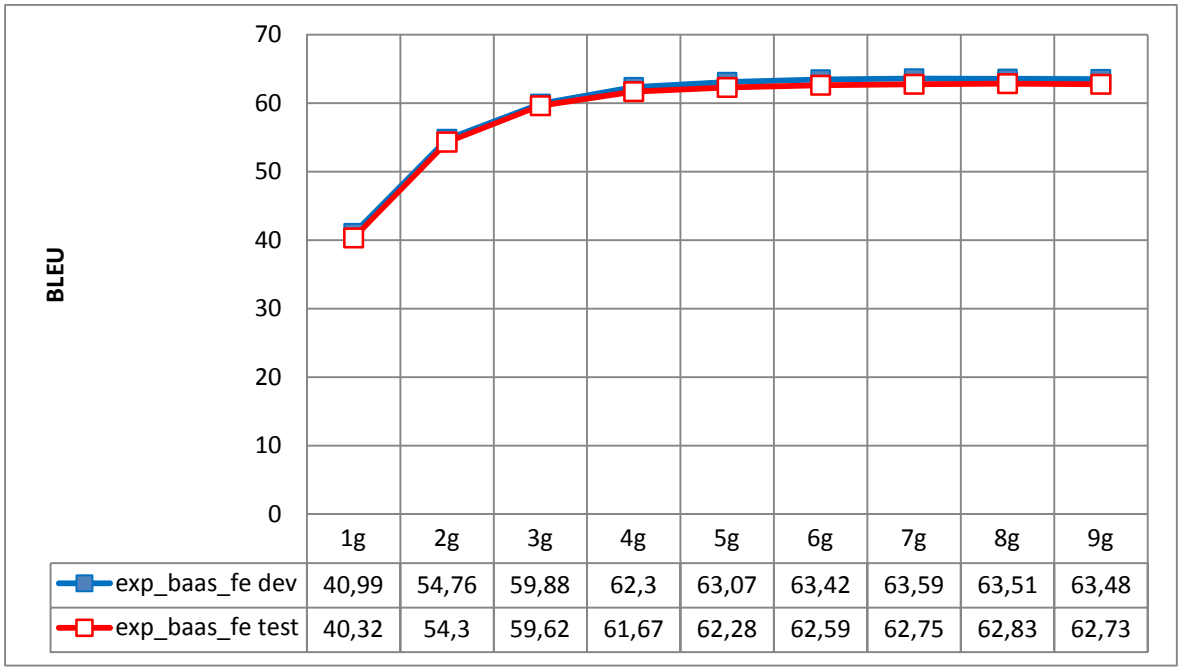
Tabel 7. Kolmanda tõlkesuuna korpuse suurus peale eeltöötlust

JRC-Acquis korpus (prantsuse-inglise tõlkesuund)

• treenimiskorpus 1 016 905 lauset
  ○ prantsuse osa - 28 633 477 sõna
  ○ inglise osa - 25 202 635 sõna
• arenduskorpus 2500 lauset
• testimiskorpus 2500 lauset

2.5.1 Baasmudel kolmandale tõlkesuunale

Sarnaselt eelnevale kahele tõlkesuunale (alampeatükk „2.4.1 Baasmudeli loomine“) peab ka kolmanda tõlkesuuna jaoks paika panema nn baassüsteemi skoori, mille põhjal oleks võimalik hinnata järgnevates eksperimentides tehtud muutusi. Selle teostamiseks sai prantsuse-inglise tõlkesuuna puhul katsetatud primaarsete keelemudelite suurust 1-gramm mudelist kuni 9-gramm mudelini. Käesolevas töös on selle baassüsteemi eksperimendi tähiseks *„exp_baas_fe*“. Allolev joonis 26 toob ära saadud tulemused.



	1g	2g	3g	4g	5g	6g	7g	8g	9g
exp_baas_fe dev	40,99	54,76	59,88	62,3	63,07	63,42	63,59	63,51	63,48
exp_baas_fe test	40,32	54,3	59,62	61,67	62,28	62,59	62,75	62,83	62,73

Joonis 26. Prantsuse-inglise baassüsteemi tulemused 1-gramm - 9-gramm

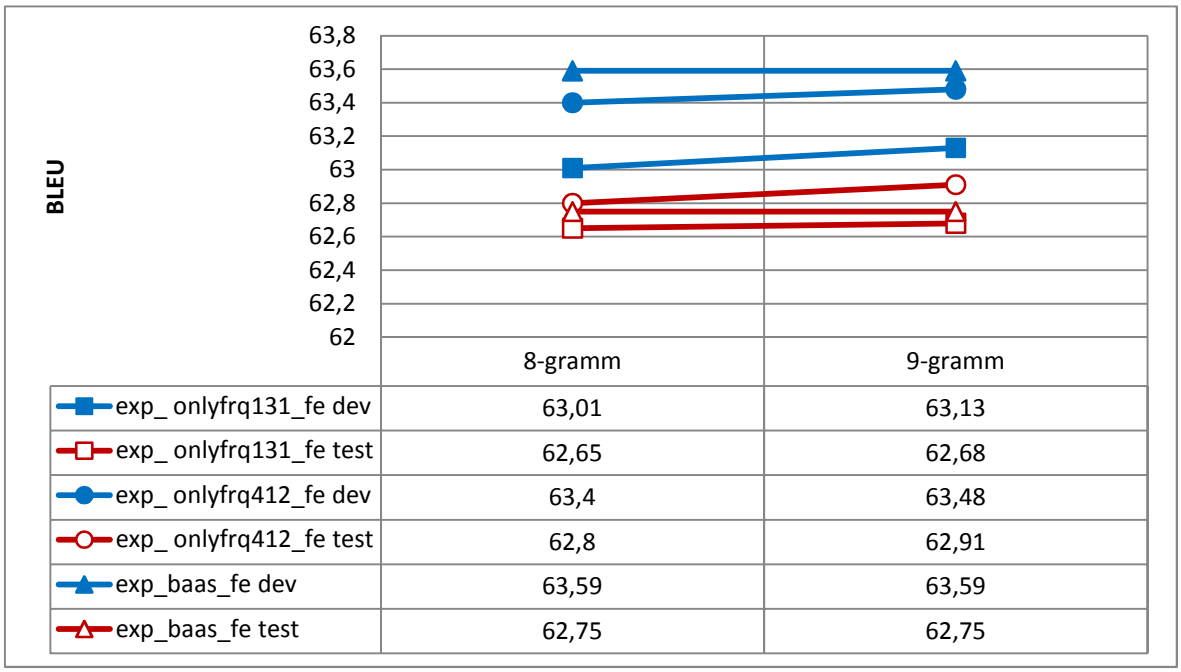
Nagu tulemustest näha, osutus prantsuse-inglise tõlkesuuna korral (sarnaselt eesti-inglise tõlkesuunale) optimaalseks 7-gramm suurusega primaarne keelemudel. Selle mudeliga saadud BLEU skoorid (vastavalt siis 63,59 treeningkorpusel ja 62,75 testimiskorpusel) on ühtlasi ka vastavateks baasmudeli skoorideks, millega võrreldakse antud tõlkesuuna järgnevaid eksperimente.

2.5.2 Esinemissageduste alusel loodud teine faktor (kolmas tõlkesuund)

Esimeseks põhieksperimendiks prantsuse-inglise tõlkesuuna jaoks sai tehtud inglise-prantsuse tõlkesuuna korral parimaid tulemusi andnud eksperiment, kus sõnad jagati kahte hulka nende esinemissageduste järgi. Nagu eelneva kahe tõlkesuuna puhul, sai ka antud juhul valitud piir sagedaste ja mittesagedaste sõnade vahel heuristilisel teel sagedus-

nimekirja uurides. Välja sai valitud kaks erineva pikkusega sagedaste sõnade nimekirja, esimene 131 ja teine 412 sõna pikk.

Käesolevas töös on selle eksperimendi tähiseks „*exp_onlyfrq[n]_fe*“. Saadud tulemused võrdluses baassüsteemi tulemustega on näha alloleval joonisel (joonis 27).



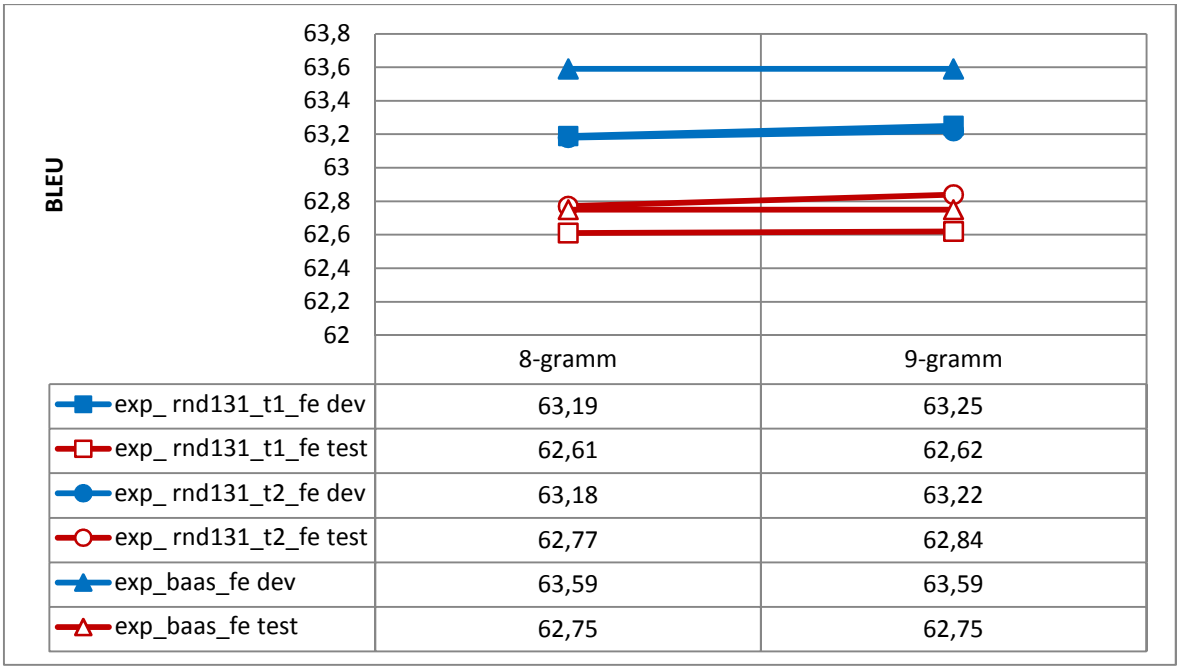
Joonis 27. Prantsuse-inglise *exp_onlyfrq131* ja *exp_onlyfrq412* tulemused võrdluses baassüsteemi tulemustega

Nagu tulemustest nähtub, on tulemused alla baassüsteemi skooride, ning seega sarnanevad eesti-inglise tõlkesuunal saadud tulemustega.

2.5.3 Juhuslikkuse alusel loodud teine faktor (kolmas tõlkesuund)

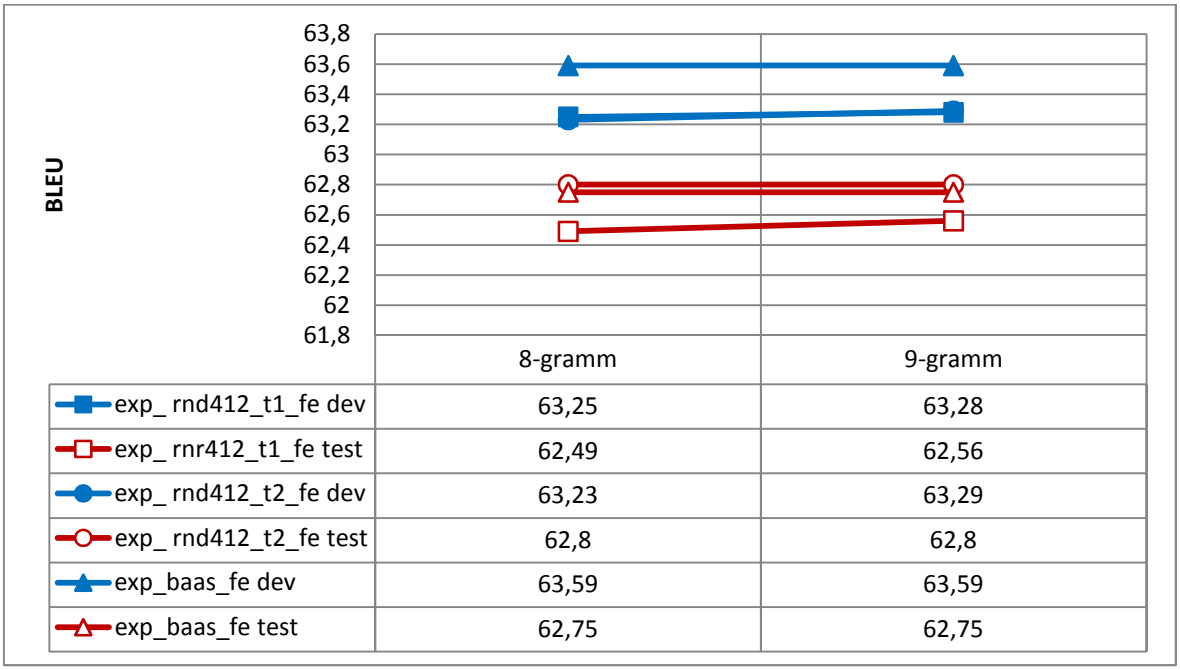
Teiseks põhieksperimendiks kolmanda tõlkesuuna jaoks sai uuesti tehtud kontroll-eksperiment, kus olid sõnad jagatud kahte hulka juhuslikkuse alusel. Käesolevas töös on selle eksperimendi tähiseks „*exp_rnd[n]_t[jrk]_fe*“. Saadud tulemused võrdluses baas-tulemuste ning eelneva eksperimendi parimate tulemustega on näha allolevatel joonistel (joonis 28 ja joonis 29).

Nagu joonistelt näha, siis antud eksperimendi skoorid jäävad alla baassüsteemi skooride, ning seega sarnanevad esimesele (eesti-inglise) tõlkesuunale.



	8-gramm	9-gramm
exp_ rnd131_t1_fe dev	63,19	63,25
exp_ rnd131_t1_fe test	62,61	62,62
exp_ rnd131_t2_fe dev	63,18	63,22
exp_ rnd131_t2_fe test	62,77	62,84
exp_baas_fe dev	63,59	63,59
exp_baas_fe test	62,75	62,75

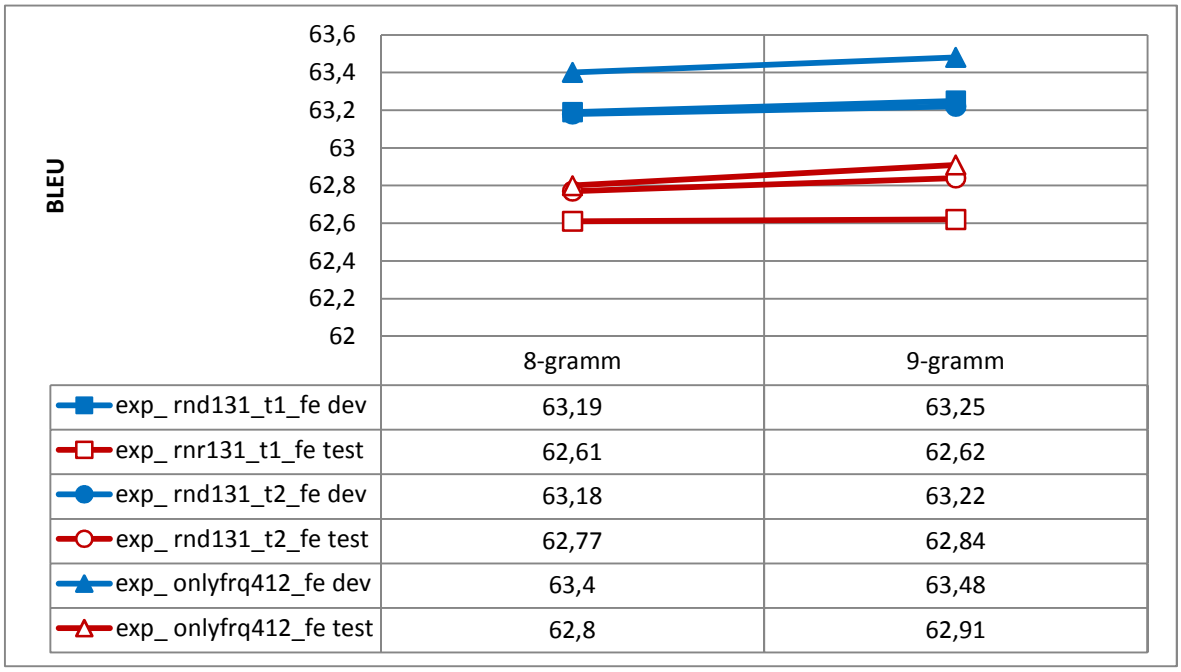
Joonis 28. Prantsuse-inglise *exp_rnd131_t1* ja *exp_rnd131_t2* tulemused võrdluses baassüsteemi tulemustega



	8-gramm	9-gramm
exp_ rnd412_t1_fe dev	63,25	63,28
exp_ rnr412_t1_fe test	62,49	62,56
exp_ rnd412_t2_fe dev	63,23	63,29
exp_ rnd412_t2_fe test	62,8	62,8
exp_baas_fe dev	63,59	63,59
exp_baas_fe test	62,75	62,75

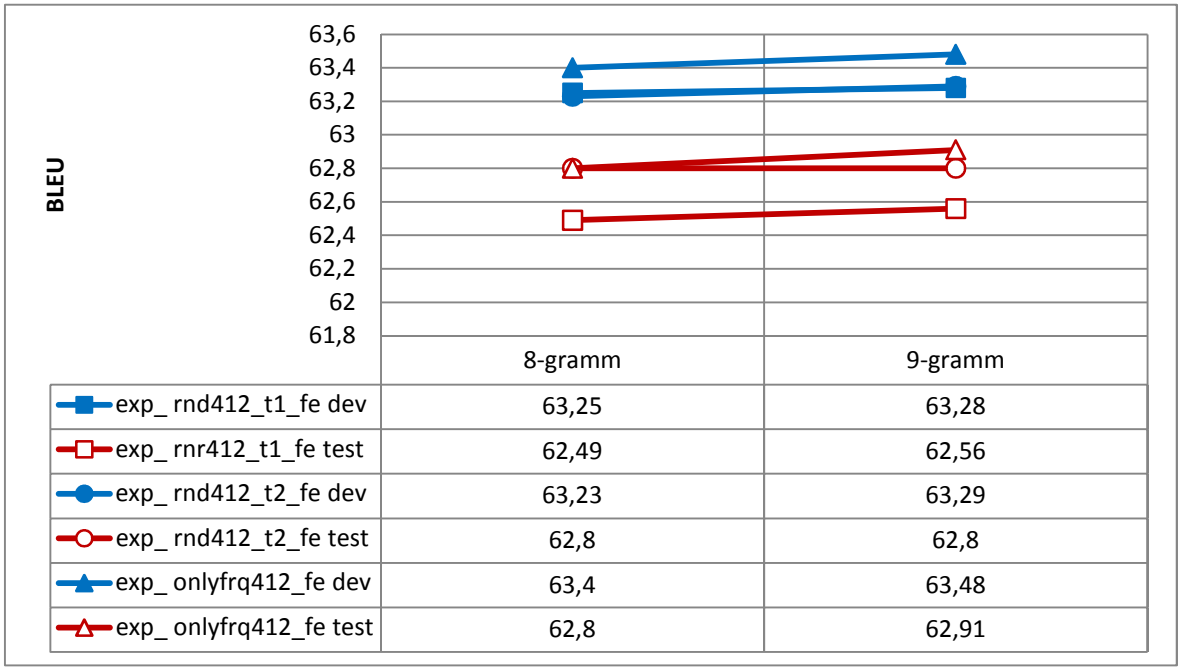
Joonis 29. Prantsuse-inglise *exp_rnd412_t1* ja *exp_rnd412_t2* tulemused võrdluses baassüsteemi tulemustega

Võrreldes antud eksperimendi tulemusi eelneva eksperimendiga, siis on näha, et tulemused on suhteliselt lähedaste skooridega, ning et sageduse alusel loodud teise faktori strateegia ei anna võrreldes juhuslikkuse kasutamisega arvestatavaid BLEU skooride paranemisi.



	8-gramm	9-gramm
exp_ rnd131_t1_fe dev	63,19	63,25
exp_ rnr131_t1_fe test	62,61	62,62
exp_ rnd131_t2_fe dev	63,18	63,22
exp_ rnd131_t2_fe test	62,77	62,84
exp_ onlyfrq412_fe dev	63,4	63,48
exp_ onlyfrq412_fe test	62,8	62,91

Joonis 30. Prantsuse-inglise *exp_rnd131_t1* ja *exp_rnd131_t2* tulemused võrdluses eksperimendi *onlyfrq412* tulemustega



	8-gramm	9-gramm
exp_ rnd412_t1_fe dev	63,25	63,28
exp_ rnr412_t1_fe test	62,49	62,56
exp_ rnd412_t2_fe dev	63,23	63,29
exp_ rnd412_t2_fe test	62,8	62,8
exp_ onlyfrq412_fe dev	63,4	63,48
exp_ onlyfrq412_fe test	62,8	62,91

Joonis 31. Prantsuse-inglise *exp_rnd412_t1* ja *exp_rnd412_t2* tulemused võrdluses eksperimendi *onlyfrq412* tulemustega

Eksperimendi *exp_rnd* kokkuvõtteks võib öelda, et prantsuse-inglise tõlkesuuna puhul annab juhuslikkusel alusel loodud teine faktor peaaegu sama häid keelemudeleid kui

sageduse alusel loodu puhul. Seega sarnanevad kolmanda tõlkesuuna tulemused eesti-inglise tõlkesuuna tulemustele.

2.6 Saadud tulemuste analüüs

Kahel esimesel, eesti-inglise ja inglise-prantsuse tõlkesuunal keelemudelite genereerimisega läbiviidud eksperimendid andsid samade strateegiate puhul erinevaid tulemusi. Eesti-inglise tõlkesuunal JRC-Acquis korpuse korral ei saadud tõlkekvaliteedi paranemist võrreldes baassüsteemi tulemustega, samas erinevate strateegiate vahelised erinevused BLEU skoorides olid nähtavad. Teise, inglise-prantsuse tõlkesuuna Europarl korpuse korral olid nähtavad vahed nii erinevate strateegiate BLEU skoorides kui ka ületasid parima strateegiaga saavutatud skoorid vastava tõlkesuuna baassüsteemi skoore. Kolmanda tõlkesuuna eksperimendid, kus kasutati esimese suunaga sama korpust ja sihtkeelt, näitasid sarnaseid tulemusi esimesele tõlkesuunale. Seega võib öelda, et antud töös katsetatud strateegiate erinev käitumine kasutatud tõlkesuundadel oli tingitud kas inglise keelest kui sihtkeelest, kasutatud JRC-Acquis korpuse valdkonnast või mõlemast. Selle väljaselgitamine, mis täpselt erinevused põhjustas, nõuaks lisakatsete läbiviimist. Võimalikud lisaeksperimendid oleks näiteks võttes kasutusele prantsuse-inglise tõlkesuuna Europarl korpusel või inglise-prantsuse tõlkesuuna JRC-Acquis korpusel või mõlemad. Käesoleva töö ajaliste piirangute tõttu nimetatud eksperimendid selle töö sisse ei mahtunud, aga need katsed on autoril plaanis tulevaste tööde käigus läbi viia.

Katsetatud neljast teise faktori põhjal sekundaarse keelemudeli loomise strateegiast andis inglise-prantsuse tõlkesuuna puhul kõige paremad tulemusi eksperiment *exp_onlyfrq*, mille korral koosnes teine faktor treeningkorpuses sagedasti esinevate sõnade puhul sõnavormidest, ülejäänud sõnade puhul sõnaliikidest. Peaaegu sama hea tulemuse andis ka eksperiment *exp_frqandpos*, mille puhul oli teises faktoris sõnavorm ka lisaks sagedastele sõnadele kõikide suletud klassi sõnade puhul. Eksperimendid *exp_wfpos* (teises faktoris suletud klassi sõnadel sõnavorm, avatud klassi omadel - sõnaliik) ja *exp_pos* (teine faktor koosnes ainult sõnaliikidest) andsid mõlemad kahest eelnevast halvemaid tulemusi. Kõige nõrgemad tulemused olidki *exp_pos* eksperimendil, mille tulemused olid isegi alla antud suuna baassüsteemi tulemuste. Eesti-inglise ja prantsuse-inglise tõlkesuundade puhul käitusid eelnevalt kirjeldatud strateegiad enam-vähem analoogselt, aga samas olid kõik alla baassüsteemi tulemuste, ning parimad tulemused olid väga sarnased juhuslikkuse alusel

loodud sekundaarse keelemudeliga eksperimentidele. Nende viimaste tulemuste najal on võimalik oletada, et antud juhtudel eelistas masintõlkesüsteem võimalikult suurt sõnavormidel loodud keelemudelit (juhuslikkuse alusel loodud sekundaarne keelemudel sisaldas väga palju sõnavorme) üldisema, sõnaliikidel loodud sekundaarse keelemudeli kasutamisele.

Eksperimendi *exp_onlyfrq* parimad tulemused on selgitatavad sellega, et mingi kindla valdkonna korpustes pole kõige sagedasemad sõnad ainult suletud klassi sõnad, vaid seal on ka teisi sellele korpusele iseloomulikke ja väga tihti esinevaid sõnu[13], ning antud strateegia võttis seda paremini arvesse ning andis natuke paremaid tulemusi kui *exp_frqandpos.*

Katsed erinevate keelemudelite suurustega andsid huvitavaid tulemusi. Tähtsamaks tõdemuseks oli see, et kuigi enamasti kasutatakse primaarsete keelemudelite loomisel masintõlkerakenduste poolt vaikimisi pakutud 3-gramm keelemudelit (selle tingis algselt tõenäoliselt piiratud mälumaht), siis ei pruugi see olla kaugeltki mitte optimaalne suurus, vaid tuleks kindlasti katsetada erinevaid suurusi. Antud töös kasutatud tõlkesuundade korral tuli esimese ja kolmanda tõlkesuuna puhul optimaalseks keelemudeli suuruseks 7-gramm mudel, teise tõlkesuuna puhul 4-gramm mudel.

Ka sekundaarse keelemudeli loomisel katsetati erineva suurusega keelemudeleid. Kuigi erinevate tõlkesuundade korral testitud keelemudelite n-gramm suurused erinesid, kehtis üldiselt trend, et suurem mudel andis paremaid tulemusi.

Samade eksperimentide käigus testimiskorpusel saadud tulemused toetasid arenduskorpuse tulemustel tehtud järeldusi.

[13] Osaline sagedusnimekiri toodud lisa 2 all.

Kokkuvõte

Käesoleva töö käigus katsetati võimalust kasutada statistilises masintõlkes morfoloogilist lisainformatsiooni sõnaliikide näol. Kasutades TreeTagger sõnaliikide märgendajat, lisati eesti-inglise, inglise-prantsuse ja prantsuse-inglise tõlkesuundade sihtkeelte korpustesse sõnaliikide tähistused, ning kasutati seda informatsiooni statistilise masintõlke jaoks lisakeelemudeli loomiseks. Lisamudeli loomise põhieesmärgiks oli uurida võimalusi kombineerides erinevatel andmetel (sõnavormidel, sõnaliikidel või nende kombinatsioonidel) loodud keelemudelit, ja saada keeleliselt korrektsemaid sihtkeelseid tõlkeid.

Läbiviidud eksperimentide tulemused olid erinevatel tõlkesuundadel erinevad. Töös katsetatud strateegiad ei andnud võrreldes baassüsteemiga positiivseid tulemusi eesti-inglise ja prantsuse-inglise tõlkesuundadel, samas paranes tõlkekvaliteet inglise-prantsuse tõlkesuuna puhul. Parimaid tulemusi antud suuna puhul andis eksperiment, kus sekundaarne keelemudel loodi sõnavormidest ja sõnaliikidest koosneva faktori põhjal, kusjuures sõnavormide ja sõnaliikide omavaheline vahekord sõltus sõnade esinemissagedustest treeningkorpuses. Sagedased sõnad olid antud faktoris esindatud sõnavormidena, vähem sagedased sõnaliikidena. Primaarse keelemudelina kasutati ainult sõnavormidel loodud keelemudelit. Parimad saadud tulemused kirjeldatud strateegiaga arendus- ja testimiskorpusel inglise-prantsuse tõlkesuuna korral olid vastavalt BLEU skoorid 33,37 ja 33,11 üle vastavate baassüsteemi skooride 32,59 ja 32,76.

Üldiselt ostutus antud töös katsetatud sekundaarsete keelemudelite kasutamine statistilises masintõlkes edukaks, samas sõltus saadud skooride paranemine suuresti kindla sihtkeele ja kasutatud korpuse valdkonna kombinatsioonist, nõudes lisakatseid tulevastes töödes.

Teiseks tähtsaks järelduseks oli see, et nii primaarsete- kui lisakeelemudelite loomisel ei tuleks piirduda masintõlkerakenduste poolt vaikimisi pakutud mudelite suurustega, vaid tuleks kindlasti uue tõlkesuuna lisamise korral katsetada ka teistsuguseid keelemudeli suurusi.

Language model based improvements in statistical machine translation

Master thesis

Harri Kirik

Abstract

The task of this thesis was to apply morphological information in statistical factored machine translation for the purpose of creating additional language models. The used morphological information consisted of annotation with POS tags and it was done by using the TreeTragger annotation application. All of the experiments in this work were carried out by using Moses statistical machine translation framework and the results were evaluated by using BLEU evaluation metric.

For the translations three different translation directions and two different corpora were used: Estonian-English (JRC-Acquis corpus), English-French (Europarl corpus) and French-English (JRC-Acquis corpus).

To goal of experiments in this work was to create and use a secondary language model in addition to primary model in statistical machine translation. While the primary language model was created on the word forms of the destination language (1st factor in the destination language corpus), the secondary language model was created on the 2nd factor containing only pos tags or containing a mixture of pos tags and word forms. The abovementioned 2nd factor was generated using different strategies, like for example taking account of the word classis and/or the frequency of the word in the training corpora.

From the experiments it was determined that the Estonian-English and French-English translation directions the results did not give any improvements on the baseline systems for these directions. But the English-French translation direction got a better translation performance when a second language model was added. And the best strategy for this second language model was the one where more frequent words were represented as word forms and less frequent words as POS tags.

Overall it was found that the efficiently of using a secondary language model made on the POS tags or on the mixture of POS tags and word forms seemed to depend on the combination of destination language and the corpora used.

Kasutatud kirjandus

Brown, P. F., Pietra, V. J., Pietra, S. A., & Mercer, R. L. (1993). The mathematics of statistical machine translation. *Computational Linguistics, 19*(2), lk 263-313.

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. *11th Conference of the European Chapter of the Association for Computational Linguistics* (lk 249–256). EACL.

Fishel, M., & Kirik, H. (2010). Linguistically Motivated Unsupervised Segmentation for Machine Translation. *LREC.* Valetta, Malta.

Kirik, H. (2008). Juhendamata morfoloogia statistilises masintõlkes. Tartu: University of Tartu, Dept. of Computer Science.

Kirik, H., & Fishel, M. (2008). Modelling Linguistic Phenomena with Unsupervised Morphology for Improving Statistical Machine Translation. *SLTC'08 Workshop on Unsupervised Methods in NLP.* Stockholm, Sweden.

Koehn, P. (2004). Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. *AMTA 2004.*

Koehn, P. (2005). *Europarl: A Parallel Corpus for Statistical Machine Translation.* Phuket Island, Thailand: MT Summit.

Koehn, P. (2010). *Statistical Machine Translation.* New York: Cambridge University Press.

Koehn, P., Federico, M., Shen, W., Bertoldi, N., Bojar, O., Callison-Burch, C., et al. (2006). *Open Source Toolkit for Statistical Machine Translation: Factored Translation Models and Confusion Network Decoding.* Johns Hopkins University, Center for Speech and Language Processing.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical Phrase-Based Translation. *HLT-NAACL.* Edmonton, Canada.

Och, F. J., & Hermann, N. (March 2003. a.). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics, 29*, lk 19-51.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics.*

Schmid, H. (1994). *Probablistic Part-of-Speech Tagging Using Decision Trees.* Stuttgart.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*(30), 50-64.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., et al. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006).* Genoa, Italy: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006).

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. *Proc. Intl. Conf. Spoken Language Processing.* Denver, Colorado.

Lisad

Käesoleval tööl on järgnevad lisad:

- Lisa 1 – kirjeldab töö käigus läbiviidud eksperimentide teostamiseks loodud skripte
- Lisa 2 – toob näiteid töö käigus sihtkorpustest genereeritud unigramm sagedus-nimekirjade kohta
- Lisa 3 – arhiiv CD plaadil, mis sisaldab lisa 1 all kirjeldatud faile

Lisa 1. Loodud skriptid

Tabel 8 kirjeldab tähtsamaid eksperimentide käigus loodud skripte ja nende põhiülesandeid. Mainitud skripte kasutati korpuse töötlemiseks ja sinna teise faktori loomiseks, seega on need vajalikud juhul, kui on soov antud töös läbiviidud eksperimente korrata.
Ülejäänud, antud tabelis mitte esinevad skriptid (ja mõned abifailid) on lühidalt ära nimetatud tabelile järgnevas loetelus.

Tabel 8. Korpuste töötluseks kasutatud skriptid

Skript, argumentide loend, lisainfo
generateOpenAndClosedStatisticsTagger.py closed_tag_list input_filename Argumendid: • closed_tag_list – nimekiri suletud POS märgenditest, igal real üks märgend, loetakse kuni tühikuni, peale seda kommentaar. Näide: „*CC coordinating conjunction, example and*“. • input_filename – sagedusnimekirja fail, mida analüüsida. Info: • Skript kuvab väljundisse statistikat sisendfailis leiduvate sõnade kohta. Antud skripti kasutati unigrammide esinemissageduste uurimisel ja sõnavormide ning sõnaliikide jaotamisel teise keelemudeli eksperimentides.
openAndClosedTaggerPos.py closed_tag_list tagged_corpora output_corpora Argumendid: • closed_tag_list – nimekiri suletud POS märgenditest, igal real üks märgend, loetakse kuni tühikuni, peale seda kommentaar. Näide: „*CC coordinating conjunction, example and*“. • tagged_corpora – faktoriseeritud kujul treenimise-, arendus- või testikorpus, mis on eelnevas failis kirjeldatud POS märgenditega märgendatud. Näide: „*I|PRP live|VBP here|RB .|.*“. • output_corpora – väljundfaili asukoht, kuhu modifitseeritud korpus kirjutada. Info: • Vaatab läbi sisendfaili ja muudab kõikide suletud klassi sõnade teise faktori sõnaliigist sõnavormiks. Avatud klassi sõnad jätab puutumata. Näide: „*and|CC*“ -> „*and|and*“

openAndClosedTaggerFrq.py wordlist tagged_corpora output_corpora [-v]

Argumendid:

- wordlist – nimekiri sõnadest kujul „sõnavorm|sõnaliik[tab]sagedus“ Näide: „*the|DT 2157276*“.
- tagged_corpora – faktoriseeritud kujul treenimise-, arendus- või testikorpus, mis on POS märgenditega märgendatud. Näide: „*I|PRP live|VBP here|RB* .|.“.
- output_corpora – väljundfaili asukoht, kuhu modifitseeritud korpus kirjutada.

Info:

- Vaatab läbi sisendfaili ja muudab kõikide etteantud nimekirjas esinevate sõnade teise faktori sõnaliigist sõnavormiks. Ülejäänud sõnad jätab puutumata.

generateRndWordList.py template_wordlist full_wordlist output_wordlist

Argumendid:

- template_wordlist – mall, mida kasutatakse sageduse arvutamiseks.
- full_wordlist – treeningkorpuses leiduvate unigrammide nimekiri koos esinemissagedustega.
- output_wordlist - väljundfaili asukoht, kuhu loodud nimekiri kirjutada.

Info:

- Võimaldab genereerida skriptile *openAndClosedTaggerFrq.py* juhusliku sisendit, mis sisaldab sama suure esinemissagedusega hulga unigramme kui etteantud mall.

Lisaks kirjeldatud skriptidele sai töö läbiviimiseks kirjutatud veel mitmeid skripte ja abifaile, need on järgnevad:

1. **decAll.sh** ja **decAllBaseline.sh** – tõlkehüpoteeside genereerimise lihtsustamiseks loodud skriptid.
2. **lines_mn.py** – katkenud eksperimentide jätkamiseks poolelijäänud kohast.
3. **evalAll.sh** ja **evalAllBaseline.sh** – arendus- ja testimiskorpuste hüpoteeside hindamise lihtsustamiseks loodud skriptid.
4. **makExtractedFactorFiles.sh** – faktoriseeritud korpusefailide eraldi alamfailidesse kirjutamiseks mõeldud abiskript.
5. **makLm.sh** – korpuse 2. faktori keelemudelite loomise lihtsustamiseks loodud skript.

6. **makModelLinks.en.sh**, **makModelLinks.ef.sh** ja **makModelLinks.fe.sh** – tõlkemudeli osade korduvkasutuseks vajalikud abivahendid.
7. **showOpenAndClosed.py** – sagedusnimekirja paremaks visuaalseks kontrollimiseks loodud skript.
8. **generateOpenAndClosedStatistics.py** – sarnane vahend eelnevalt kirjeldatud skriptile *generateOpenAndClosedStatisticsTagger.py*, aga mõeldud Connexor[14] märgendaja XML väljundile.
9. **makeListWithAnalysis.py** – abivahend Connexor märgendajaga märgendatud XML failist unigrammide nimekirja loomiseks.
10. **wordAndPosFactors.py** – Connexor märgendajaga töödeldud korpuse viimiseks vajalikule faktorkujule (XML dokumendist tekstikorpuse tagasi).
11. **wordAndWordClassFactors.py** – samasugune abivahend nagu eelnevalt kirjeldatud *openAndClosedTaggerPos.py*, aga Connexor märgendaja XML väljundile.
12. **wordAndWordFrequencyFactors.py** - sama vahend nagu eelnevalt kirjeldatud openAndClosedTaggerFrq.py, aga Connexor märgendaja XML väljundile.
13. **cleanConexorOutput.py**, **cleanConexorTextOutput.py**, **outputWordsWithSpaces.py** – vahendid Connexor programmi väljundi analüüsiks ja teisendamiseks.
14. **tagger_closed_tags_en.txt** ja **tagger_closed_tags_fr.txt** – abifailid, mis toovad ära meie eksperimentide jaoks suletud klassi loetavate sõnaliikide märgendite loetelu.
15. **marked.sorted.count.1gram.en, marked.sorted.count.1gram.fr ja marked.sorted.count.1gram.fr_en** failid – abifailid, sisaldavad treeningkorpuses sisaldunud unigrammide täielikke nimekirju.

[14] Connexor rakendust sai kasutatud töö varasemas etapis, aga sellega märgendatud eksperimendid ei jõudnud lõplikku töösse. Täpsem info Connexor rakenduse kohta lehel: http://www.connexor.eu

Lisa 2. Sagedusnimekirjade näiteid

Järgnevalt on ära toodud näited tõlkesuundade sihtkeelsetel treeningkorpustel koostatud unigrammnimekirjade kohta. Näidete eesmärgiks on illustreerida olukorda, kus töö käigus kasutatud korpuste kõige sagedasemad sõnad polnud ainult suletud klassi sõnad, vaid sisaldasid mitmeid väga sagedasi avatud klassi sõnu. Enamasti olid need sagedased avatud klassi sõnad spetsiifilised antud valdkonnale, mille aluselt oli treeningkorpus koostatud.

Kuna töös kasutatud kolme tõlkesuuna puhul olid esimese ja kolmanda suuna korral sihtkeeled samad, siis tuuakse katkendid ainult ingliskeelsel ja prantsuskeelsel korpusel koostatud sagedusnimekirjadest.

Tabel 9 toob ära viiskümmend sagedasemat unigrammi JRC-Acquis korpuse ingliskeelses (summaarne kogusagedus 27 911 130) ja Europarl korpuse prantsuskeelses osas (summaarne kogusagedus 40 803 011). Suletud klassi arvatud unigrammid on märgitud märgendiga „SK“, kõik ilma vastava märgendita unigrammid arvati avatud klassi sõnade hulka.

Tabel 9. Viiskümmend sagedasemat unigrammi inglise- ja prantsuskeelsetes treeningkorpustes

	Inglise korpus (JRC-Acquis)				Prantsuse korpus (Europarl)		
1	the|DT	2 157 276	SK		,|PUN	1 808 662	SK
2	of|IN	1 192 616	SK		de|PRP	1 796 107	SK
3	,|,	961 367	SK		'|PUN	1 505 276	SK
4	.|SENT	839 552	SK		.|SENT	1 318 532	SK
5	to|TO	714 963	SK		la|DET:ART	1 165 391	SK
6	in|IN	670 969	SK		et|KON	808 678	SK
7	and|CC	620 785	SK		le|DET:ART	766 766	SK
8	)|)	501 466	SK		à|PRP	717 285	SK
9	(|(	497 314	SK		les|DET:ART	677 696	SK
10	for|IN	329 280	SK		des|PRP:det	642 622	SK
11	a|DT	310 929	SK		que|KON	461 765	SK
12	be|VB	270 783			en|PRP	443 145	SK
13	/|SYM	247 517	SK		l|NOM	434 480	
14	on|IN	227 039	SK		nous|PRO:PER	399 522	SK
15	article|NP	212 623			est|VER:pres	385 160	
16	by|IN	205 358	SK		-|PUN	359 166	SK
17	shall|MD	203 202			une|DET:ART	350 375	SK
18	-|:	202 292	SK		du|PRP:det	339 185	SK
19	or|CC	165 070	SK		dans|PRP	322 847	SK
20	is|VBZ	164 781			un|DET:ART	317 478	SK
21	with|IN	160 864	SK		’|NOM	309 521	
22	this|DT	150 917	SK		il|PRO:PER	305 146	SK

23	1|CD	130 770			qui|PRO:REL	293 891	SK
24	commission|NP	123 622			ce|PRO:DEM	276 383	SK
25	member|NP	123 211			pour|PRP	272 386	SK
26	:|:	122 011	SK		je|PRO:PER	269 748	SK
27	;|:	118 094	SK		’|ADJ	236 809	
28	that|IN	113 431	SK		pas|ADV	229 982	
29	which|WDT	104 695	SK		au|PRP:det	226 641	SK
30	community|NP	101 167			a|VER:pres	217 156	
31	from|IN	97 536	SK		sur|PRP	214 816	SK
32	ec|NP	94 036			d|ADJ	205 644	
33	2|CD	93 845			ne|ADV	198 632	
34	are|VBP	92 724			par|PRP	184 190	SK
35	regulation|NP	89 968			plus|ADV	150 564	
36	as|IN	88 911	SK		commission|NOM	149 473	
37	not|RB	88 901			cette|PRO:DEM	137 709	SK
38	it|PP	88 606	SK		l|VER:pper	136 296	
39	states|NPS	85 250			d|VER:pper	121 635	
40	no|NP	85 110			aux|PRP:det	112 891	SK
41	european|NP	73 329			mais|KON	107 314	SK
42	an|DT	70 282	SK		sont|VER:pres	104 764	
43	at|IN	65 190	SK		être|VER:infi	104 246	
44	council|NP	64 467			que|PRO:REL	102 740	SK
45	should|MD	63 423			l|ADJ	100 210	
46	its|PP$	63 028			avec|PRP	97 486	SK
47	has|VBZ	61 743			président|NOM	95 611	
48	regulation|NN	59 564			européenne|ADJ	93 531	
49	as|RB	58 046			ces|PRO:DEM	90 593	SK
50	2|LS	57 354			union|NOM	90 191	

Nimekirjade automaatseks genereerimiseks kasutati järgnevaid samme:

1. Kasutades keelemudeli loomiseks mõeldud rakendust SRILM genereeriti nimekiri (juba sõna-liikidega märgendatud) treeningkorpuses sisalduvatest unigrammidest ja nende sagedustest:
    - Näide: *ngram-count -order 1 -text train.en -write count.1gram.en*
2. Eelneva sammu juures saadud nimekiri sorteeriti kahanevalt sageduste järgi:
    - Näide: *sort -nr -k2 count.1gram.en > sorted.count.1gram.en*