ARVI HURSKAINEN

Tiedonhaku Raamatusta mullistuu

JOHDANTO

Teologiaa opiskelleet ja varmaan myös monet maallikot tuntevat Vilho Vuorelan toimittaman vuodelta 1962 peräisin olevan kaksiosaisen Raamatun hakusanakirjan.[1] Vuoden 1938 raamatunkäännökseen pohjautuva hakuteos on ollut vuosikymmenten ajan korvaamaton apuväline Raamatusta kiinnostuneille henkilöille, ja on varmaan sitä edelleenkin. Digitekniikka on kuitenkin tuonut mukanaan uudenlaisia tiedonhakumahdollisuuksia. Aluksi esittelen lyhyesti nykyisin saatavilla olevia digitaalisia hakusovelluksia.

Internetissä on saatavilla muutamia raamattuhakuun soveltuvia palveluita,[2] kuten koivuniemi.com[3] ja uskonkirjat.net.[4] Jälkimmäinen tarjoaa hakupalveluja myös muista kuin suomenkielisistä raamatunkäännöksistä. Molemmat palvelut edustavat ensimmäisen polven digitaalisia hakusovelluksia. Niissä käytetään suoraa merkkijonohakua, eli hakuavaimen ja löydetyn sanan täytyy täsmätä. Hakuja voi kuitenkin suorittaa monella eri tavalla. Hakujen tekemistä tukee molemmissa palveluissa olevat hyvät käyttöohjeet.

Suomen Pipliaseuralla on käytössään uusi aiempaa edistyneempi hakujärjestelmä.[5] Hakusanalla voi löytää myös sellaiset jakeet, joissa hakusana esiintyy taivutetussa muodossa. Suomenkielisille lukijoille kyseessä on suuri edistysaskel, sillä suomessa esiintyvät monet taivutusmuodot hankaloittavat hakua. Pipliaseuran järjestelmän puutteena on kuitenkin se, että siinä ei voi muodostaa täysin yksiselitteisiä hakuavaimia. Esimerkiksi kun hakukenttään kirjoitetaan sana *tuli*, hakutuloksessa esiintyvät kaikki tuon merkkijonon sisältävät jakeet, hakuavain tummennettuna. Lisäksi hakutulokseen sisältyy verbin *tulla* taivutettuja muotoja, mutta ei substantiivin *tuli* taivutettuja muotoja. Jos Pipiliaseuran hakujärjestelmästä etsii substantiivin taivutettua muotoa *tulella*, jollainen tekstissä esiintyy ja joka ei ole verbin muoto, löytyy substantiivin muita muotoja. Lisäksi

1 Vuorela 1962a, b.
2 Näiden hakujärjestelmien vertailu on ositteessa: http://www.njas.helsinki.fi/salama/evaluation-of-four-search-systems-of-finnish-bible.pdf.
3 https://www.koivuniemi.com/raamattu.
4 https://raamattu.uskonkirjat.net/servlet/biblesite.Bible.
5 https://raamattu.fi.

hakutulokseen sisältyy myös verbin *tulla* taivutettuja muotoja. Esimerkki osoittaa, kuinka vaikeaa on toteuttaa täysin kattava ja tarkka hakujärjestelmä. Mahdotonta se ei kuitenkaan ole.

Kehittämäni hakujärjestelmä Salama poikkeaa aiemmin mainituista hakujärjestelmistä siten, että perinteisten hakuavainten lisäksi siinä on mahdollista käyttää yksiselitteisiä avaimia. Tällöin haku kohdistuu sanan analysoituun muotoon, jolloin sanan perusmuotoon liitetään myös sanaluokkatunniste. Yllä mainittu haku voidaan suorittaa siten, että käytetään hakuavainta *tuli_N* substantiivien löytämiseksi ja hakuavainta *tulla_V* verbien löytämiseksi. Hakutulos on täysin virheetön, jos analysoidun tekstin disambiguaatio on virheetön. Salama sisältää myös käyttäjäystävällisen version, jossa hakuavaimen voi kirjoittaa mihin muotoon tahansa. Järjestelmä hakee sanasta sen perusmuodon ja liittää siihen oletusarvoisen sanaluokkatunnisteen. Haku tapahtuu näin muodostetun hakuavaimen perusteella. Yllä mainitun hakutehtävän suorittamiseksi hakukenttään voi kirjoittaa esimerkiksi *tulessakaan,* tai minkä tahansa muun tämän substantiivin taivutusmuodon. Raamatun tekstistä riippumaton analysaattori muuttaa sen muotoon *tuli _N*, ja haku tapahtuu analysoidun muodon perusteella. Jos taas hakukenttään kirjoittaa *tullessakaan*, analyysin tulos on *tulla _V*, ja haku tapahtuu tämän muodon perusteella. Molemmissa tapauksissa hakutulos on virheetön.

Sanojen lukuisat taivutusmuodot vaikeuttavat tiedonhakua. Optimaalisen hakuavaimen muodostaminen on usein hankalaa. Hakutuloksesta voi puuttua hakukohteita ja usein tuloksessa on myös sellaisia osumia, joita ei oltu haettu. Hyvässä hakumenetelmässä on kaksi kriteeriä, jotka sen tulee täyttää. Haun tulee olla kattava, eli kaikki haettavat kohteet löytyvät. Toisaalta haun tulee olla tarkka, jolloin tulos sisältää vain haetut kohteet.

Moni muistaa suomen kielen tunneilla opetetut puuduttavat kielen taivutuskaavat, astevaihtelut, etu- ja takavokaalisuuden aiheuttamat muutokset taivutukseen, eri verbien subjektien ja objektien sijamuodot ynnä muut vaikeasti hahmotettavat kielen yksityiskohdat. Vaikka näiden kielen piirteiden systemaattinen kuvaaminen koetaan vaikeaksi, silti osaamme vaivattomasti ja lähes virheettömästi käyttää kieltä ja sen vivahteita.

Oma käsitykseni on, että tarkka ja kattava tiedonhakuohjelma voidaan toteuttaa ainoastaan sellaisen lähestymistavan avulla, joka sekä analysoi että disambiguoi kohdetekstin jokaisen sanan. Kohdetekstillä viittaan Raamatun tekstiin. Analyysi tarkoittaa, että jokaiselle sanamuodolle määritellään kaikki sen mahdolliset tulkinnat kontekstista riippumatta. Disambiguaatio tarkoittaa, että jokaiselle sellaiselle sanamuodolle, jolla on useampi kuin yksi tulkinta, etsitään sen oikea tulkinta kyseisessä kontekstissa. Analyysin tuloksena saadaan sanan leksikaalinen muoto, joka koostuu sanan perusmuodosta eli lemmasta ja joukosta niin sanottuja tageja, joilla kuvataan sanamuodon eri piirteitä. Tässä katsauksessa otetaan analyysituloksesta mukaan ainoastaan piirre, joka määrittelee kyseisen lemman sanaluokan.

Kun kohdetekstiä rikastetaan siten, että jokaisen sanan jäljessä on kyseisen sanan lemma yhdistettynä sanaluokkatunnukseen, tuloksena on sellainen tekstimuoto, joka mahdollistaa kattavan ja tarkan haun.

Esimerkki rikastetusta tekstimuodosta:

> 1Moos 1:1 Alussa {alku_N} loi {luoda_V} Jumala {Jumala_ERISN} taivaan {taivas_N} ja {ja_KONJ} maan {maa_N}.[6]

Vain haettavasta sanasta näytetään lemmamuoto ja sanaluokkatunnus. Kaikista muista sanoista poistetaan analyysitulos. Muutamia

suomen kielen analyysiohjelmia on jo kehitetty ja jotkut niistä ovat yleisesti käytettävissä.[7] Yleinen analyysiohjelma ei kuitenkaan sovellu sellaisenaan Raamatun tekstin analysoimiseen, sillä Raamatussa on runsaasti erikoissanastoa, varsinkin erisnimiä, jotka lisäksi taipuvat. Jos halutaan kattava ja tarkka hakujärjestelmä, analyysiohjelmaa joudutaankin aina muokkaamaan niin, että se soveltuu kohdetekstiinsä.

Tämä koskee myös disambiguaatio-ohjelmaa, joka tulee kehittää käyttäen kohdetekstiä testialustana. Tarkan disambiguaation kehittäminen on itse asiassa vielä työläämpää kuin analyysiohjelman laatiminen, sillä oikean valinnan tekeminen on joskus lähes mahdotonta. Mikäli kyseessä on rajattu ja stabiili teksti kuten Raamattu, disambiguoituun tekstiin voi kuitenkin tehdä vielä korjauksia manuaalisesti. Näin on mahdollista saada myös disambiguaatiosta täydellinen.

Tämän katsauksen varsinaisena tarkoituksena on havainnollistaa käsin tehdyn hakujärjestelmän ja tietokoneella tehdyn hakujärjestelmän välisiä keskeisiä eroja. Vertailussa käytän Vuorelan Raamatun hakusanakirjaa vuodelta 1962 ja Transtechno oy:n tuottamaa Salama-nimistä järjestelmää, johon on sijoitettu joukko erilaisia Raamatun hakujärjestelmiä. Koska Vuorelan hakuteos koskee vuosien 1933/1938 raamatunkäännöstä, käytän vertailussa Salaman hakujärjestelmää, joka on sovellettu tuohon käännökseen. Vuorela on sijoittanut Vanhan testamentin ja Uuden testamentin erillisiin niteisiin. Tämän vuoksi myös Salaman hakujärjestelmä on jaettu vastaavasti kahteen osaan.

Vertailu on tehty kolmenlaisella materiaalilla. Raamatusta on ensin tuotettu lemmalista, joka on sitten jaettu erisnimiin ja tavallisiin sanoihin. Näistä molemmista erikseen on sitten muodostettu kaksi listaa. Yhdessä listassa lemmat on järjestetty esiintymistiheyden perusteella. Toisessa listassa lemmat on sekoitettu satunnaisvalintaa käyttäen. Esiintymistiheyden perusteella voidaan löytää esimerkiksi yleisimmät sanat. Satunnaisvalinnalla saadaan lista, josta voidaan tehdä otos mistä kohtaa tahansa menettämättä objektiivisuutta. Kolmas vertailutapa on valita sellaisia sanoja, joita oletettavasti usein etsitään. Se mitä tässä viimeisessä menetelmässä menetetään objektiivisuudessa, voitetaan kiinnostavuudessa.

MANUAALISEN JA DIGITAALISEN HAKUJÄRJESTELMÄN VERTAILU

Hakujärjestelmien vertailutiedot on esitetty taulukkojen muodossa. Frekvenssitiedot on tuotettu Salaman avulla. Ne siis jo sinällään kuvaavat Salaman avulla löydettyjen esiintymien määrää, joten erillistä listaa järjestelmän kattavuudesta ei tarvita. Vuorelan hakusanakirjan tiedot on poimittu käsin Vanhasta testamentista ja Uudesta testamentista erikseen. Vain sellaiset esiintymät on otettu mukaan, joissa on mukana myös kontekstia. Pelkät viittaukset Raamatun kohtaan sen sijaan eivät ole mukana.

Raamatun yleisimmät sanat

Alla olevissa taulukoissa on kuvattu hakusanojen esiintymien määrää Raamatussa. Vasemmanpuoleisissa sarakkeissa on sanojen todellinen määrä koko Raamatussa, sekä lisäksi Vanhassa ja Uudessa testamentissa erikseen. Oikeanpuolisissa sarakkeissa on vastaavat esiintymätiedot Vuorelan sanakirjassa. Viimeisessä sarakkeessa on prosenttiosuus sanojen kokonaismäärästä.

6 Perustekstistä tuotetuista rakenteisista tekstin esitysmuodoista käytetään usein termiä ”tietokanta”. Tämä ei ole kuitenkaan tietokanta, vaan ihan tavallista tekstiä.

7 https://turkunlp.org/Turku-neural-parser-pipeline/. https://github.com/flammie/omorfi

SALAMA			Vuorelan hakusanakirja			
Kaikki	VT	UT	VT	UT	Kaikki	%
8377 herra_N	7655	722	309	139	448	5,35
7000 sanoa_V	4689	2311	0	22	22	0,31
4994 tulla_V	3539	1455	6	52	58	1,16
3707 tehdä_V	2861	846	64	91	155	4,18
3288 maa_N	2958	330	413	110	523	15,91
3145 poika_N	2778	367	123	154	277	8,81
3043 kuningas_N	2917	126	275	57	332	10,91
3019 antaa_V	2314	705	20	134	154	5,10
2399 kansa_N	2058	341	261	120	381	15,88
2088 päivä_N	1690	398	207	138	345	16,52
2104 mennä_V	1508	596	4	24	28	1,33
2050 mies_N	1665	385	131	82	213	10,39
1756 ottaa_V	1395	451	10	102	112	6,38
1698 saada_V	1189	509	0	21	21	1,24
1608 katsoa_V	1249	359	18	44	62	3,86
1582 käsi_N	1356	226	280	109	389	25,59
1529 kuulla_V	1084	445	131	113	244	15,98
1491 isä_N	1072	419	157	246	403	27,00
1471 puhua_V	1033	438	56	201	257	17,49
1410 nähdä_V	888	522	40	141	181	12,84

Taulukko 1. Raamatun 20 lähes yleisintä sanaa.

Taulukko 1 havainnollistaa, miten yleisten hakusanojen esiintymien kattava kuvaaminen painetussa kirjassa ei ole mahdollista. Minkään sanan esiintymistä ei ole sanakirjassa kuin pieni murto-osa. Lisäksi joitakin sanoja on ilmeisesti pidetty tärkeämpinä kuin toisia, mikä onkin ymmärrettävää.

Satunnaisesti valitut sanat

Seuraavaksi katsomme, kuinka kattavasti satunnaisesti valitut sanat on löydetty Vuorelan sanakirjassa. Otamme satunnaisesti järjestetyn listan alusta 20 sellaista sanaa, jotka esiintyvät Raamatussa vähintään kolme kertaa (taulukko 2).

Raamatussa harvoin esiintyvistä sanoista on Vuorelan sanakirjaan listattu suhteellisesti enemmän esiintymiä kuin yleisesti esiintyvistä sanoista. Mistään otoksen sanasta ei kuitenkaan ole löydetty kaikkia esiintymiä.

Yleisimmät erisnimet

Yleisimpien erisnimien osalta tilanne on samanlainen kuin yleissanojen osalta. Vain vähäinen osa esiintymistä on listattu (taulukko 3). Lisäksi joistakin erisnimistä on yllättävän vähän esimerkkejä. Huomiota kiinnittävät sellaiset erisnimet kuin Jeesus, Saul, Aaron, Salomo ja Joosua, joista on listattu vain pieni osa.

SALAMA			Vuorelan hakusanakirja			
Kaikki	VT	UT	VT	UT	Kaikki	%
4 verityö_N	4	0	2	0	2	50
21 harhailla_V	20	1	4	2	6	28,57
123 kallio_N	109	14	68	9	77	62,60
18 syrjä_N	18	0	0	0	0	0
10 silmänräpäys_N	9	1	3	1	4	40,00
25 aalto_N	18	7	11	6	17	68,00
5 lukittu_A	3	2	1	1	2	40,00
16 tyydyttää_V	15	1	4	1	5	31,25
3 sieni_N	0	3	0	1	1	33,33
6 käsikivi_N	5	1	3	1	4	66,67
45 maanpiiri_N	36	9	19	7	26	57,78
26 pauhata_V	23	3	6	3	9	34,61
21 kauppias_N	16	5	7	5	12	57,14
53 terve_A	9	44	6	33	39	73,58
27 asuvainen_N	17	10	0	1	1	3,70
161 viisas_A	140	21	87	20	107	66,46
75 kohottaa_V	70	5	2	5	7	9,33
7 polttouhriteuras_N	7	0	0	0	0	0
3 ovipuolisko_N	3	0	0	0	0	0
24 lukuisa_A	22	2	0	2	2	8,33

Taulukko 2. Satunnaisesti valitut sanat.

SALAMA			Vuorelan hakusanakirja			
Kaikki	VT	UT	VT	UT	Kaikki	%
4042 Jumala_ERISN	2703	1339	302, 105	360	767	18,98
1956 Israel_ERISN	1887	69	53	42	95	4,86
1137 Daavid_ERISN	1078	59	39	26	65	5,72
973 Jeesus_ERISN	0	973	0	28	28	2,88
850 Mooses_ERISN	769	81	18	39	57	6,71
817 Juuda_ERISN	779	12	29	52	81	9,91
809 Jerusalem_ERISN	669	140	77	66	143	17,68
540 Egypti_ERISN	522	18	39	12	51	9,44
517 Kristus_ERISN	0	517	0	179	179	34,49
431 Jaakob_ERISN	357	69	26	32	58	13,46
413 Saul_ERISN	404	9	8	1	9	2,18
351 Aaron_ERISN	346	5	9	5	14	3,99
302 Salomo_ERISN	290	12	1	6	7	2,32
287 Baabel_ERISN	287	0	26	0	26	9,06
283 Sebaot_ERISN	281	2	17	1	18	6,36
253 Aabraham_ERISN	175	78	4	43	47	18,58
249 Joosef_ERISN	213	36	12	15	27	10,84
248 Joosua_ERISN	246	2	1	1	2	0,81
197 Jeremia_ERISN	144	53	19	1	20	10,15
187 Jordan_ERISN	172	15	22	8	30	16,04

Taulukko 3. Yleisimmät erisnimet.

SALAMA			Vuorelan hakusanakirja			
Kaikki	VT	UT	VT	UT	Kaikki	%
2 Hagaba_ERISN	2	0	0	0	0	0
3 Maaon_ERISN	3	0	0	0	0	0
6 Trooas_ERISN	0	6	0	4	4	66,67
9 Selah_ERISN	9	0	0	0	0	0
40 Ahasja_ERISN	40	0	2	0	2	5,00
134 Iisak_ERISN	114	20	8	11	19	14,18
13 Sealtiel_ERISN	10	3	0	2	2	15,38
9 Soobal_ERISN	9	0	0	0	0	0
28 Kehat_ERISN	28	0	1	0	1	3,57
5 Barsillai_ERISN	5	0	3	0	3	60,00
3 Suubael_ERISN	3	0	0	0	0	0
3 Beetfage_ERISN	0	3	0	3	3	100
4 Uuriel_ERISN	4	0	0	0	0	0
3 Jaarib_ERISN	3	0	0	0	0	0
4 Toob_ERISN	4	0	0	0	0	0
6 Sered_ERISN	6	0	0	0	0	0
2 Behemot_ERISN	2	0	1	0	1	50,00
5 Mispe_ERISN	5	0	0	0	0	0
2 Kenat_ERISN	2	0	1	0	1	50,00
3 Besek_ERISN	3	0	0	0	0	0

Taulukko 4. Satunnaisesti valitut erisnimet.

Satunnaisesti valitut erisnimet

Satunnaismenetelmällä valittujen erisnimien osalta tilanne on varsin synkkä. Useimmista nimistä ei ole listattu yhtään esiintymää. Tosin listaan (taulukko 4) on osunut vain yksi varsin yleinen erisnimi. Listasta on poistettu sellaiset nimet, jotka esiintyvät vain kerran.

Yleisimmin haetut sanat

Käytössäni ei ole tutkittua tietoa siitä, mitä sanoja Raamatusta haetaan eniten. Tähän otokseen (taulukko 5) olen valinnut sellaisia sanoja, joita itse haen usein ja oletan muidenkin tekevän samoin.

Näiden eniten haettujen sanojen kattavuus Vuorelan sanakirjassa on aivan eri luokkaa kuin muiden sanojen osalta. Sanojen oletettu kiinnostavuus lienee saanut tekijän kiinnittämään näihin erityistä huomiota. Näistäkin sanoista vain harvojen kaikki esiintymät on listattu.

VERTAILUN JA HAKUJÄRJESTELMIEN ARVIOINTIA

Oleellinen ero Salaman hakujärjestelmän ja Vuorelan hakusanakirjan välillä on niiden kattavuudessa. Edellinen löytää Raamatusta kaikki sanan esiintymät riippumatta siitä, missä muodossa sanat ovat tekstissä ja kuinka paljon niitä on. Vuorela on käyttänyt vaihtelevia menetelmiä sekä valitessaan hakusanoja ja esimerkkejä. Vain hyvin harvoin sanan kaikki esiintymät on listattu.

Vaikkakaan Vuorelan menetelmä ei ole kattava, se on varsin tarkka. Mukana ei ole vääriä esimerkkejä. Lisäksi Vuorela on tehnyt joidenkin keskeisten sanojen osalta alajaottelua, mikä on käyttäjän kannalta hyödyllistä. Tällaisten sanojen alaluokkien erotteleminen Salaman avulla on kyllä mahdollista, mutta se vaatisi semanttisten tunnisteiden lisäämistä jär-

SALAMA			Vuorelan hakusanakirja			
Kaikki	VT	UT	VT	UT	Kaikki	%
570 synti_N	354	216	207	156	363	63,68
367 armo_N	233	134	137	91	228	62,13
500 laki_N	297	203	97	143	240	48,00
109 evankeliumi_N	0	109	0	84	84	77,06
317 vanhurskaus_N	226	91	181	78	259	81,70
17 vanhurskauttaa_V	1	16	1	16	17	100
278 vanhurskas_A+N	200	78	179	65	244	87,77
312 pelastaa_V	259	53	116	43	159	50,96
136 pelastua_V	82	54	20	46	66	48,53
113 pelastus_N	71	42	57	39	96	84,96
86 autuas_A	35	51	34	35	69	80,23
910 kuolla_V	616	294	122	81	203	22,31
219 kuolema_N	92	76	99	93	192	87,67
21 kadotus_N	1	20	1	20	21	100,00
78 tuonela_N	68	10	58	9	67	85,90
12 helvetti_N	0	12	0	8	8	66,67
35 perkele_N	0	35	0	29	29	82,86
58 saatana_N	18	40	0	28	28	48,28
722 taivas_N	448	274	169	141	310	42,94
18 paratiisi_N	14	4	7	3	10	55,56

Taulukko 5. Yleisimmin haetut sanat.

jestelmään. Nykyisessä sovelluksessa tällaisia tunnisteita ei ole.

Toinen oleellinen ero näiden kahden hakujärjestelmän välillä liittyy objektiivisuuteen. Manuaalisessa työskentelyssä esimerkkien valinta on alttiina monenlaisille häiriötekijöille. Kirjoittajan henkilökohtaiset mieltymykset voivat vaikuttaa valintaan. Myös päivittäin vaihtelevat työskentelyolosuhteet voivat vaikuttaa siihen, kuinka tärkeänä mitäkin hakusanaa pidetään. Lisäksi julkaisulle määritelty maksimikoko asettaa rajoituksia. Tällöin helppo tapa pysyä annetuissa rajoissa on jättää hakusanoja pois.

Digitaalisesta tiedonhausta voi väittää, että se on objektiivista, koska edellä kuvattuja valintoja ei tarvitse tehdä. On eri asia arvioida, onko tällainen täysin kattava tiedonhaku käyttäjän kannalta aina järkevää. Jos haun tuloksena on tuhansia esiintymiä, niistä tarvittavan tiedon seulominen voi olla työlästä.

Onneksi Salaman hakujärjestelmällä tietoa voi hakea myös monella muulla tavalla kuin käyttäen sanan lemmaa hakuavaimena. Haku voidaan kohdistaa myös alkuperäiseen tekstiin, josta voi hakea taivutettujen sanojen tai sanan osien perusteella. Lisäksi voi määritellä vain sanan alkuosan tai loppuosan. Hakujärjestelmä mahdollistaa myös useamman sanan perusteella tapahtuvan haun. Esimerkiksi operaattorilla TAI voidaan tehdä hakuja vaihtoehtoisten avainten perusteella, kun taas operaattorilla JA voidaan hakea esiintymiä, joissa on vähintään kaksi hakuavainta vastaavaa sanaa.

Salaman hakujärjestelmään on myös liitetty mahdollisuus hakea esiintymiä kahden tai kolmen perättäisen sanan avulla. Näitä esiintymiä

voi hakea kahdella tavalla. Yksi tapa on hakea pintamuodon mukaan, jolloin hakuavaimen tulee olla täsmälleen sen muotoinen kuin se on tekstissä. Toinen tapa on käyttää hakuavaimena sanojen lemmamuotoja, jolloin hakuavaimessa tulee olla lemmaa tarkoittava tunniste. Lemmamuotoja käytettäessä löytyvät sanojen esiintymät niiden pintamuodosta riippumatta.

Digitaalisen hakujärjestelmän etuna on mahdollisuus kopioida hakutuloksesta halutut esiintymät ja siirtää ne omaan dokumenttiin. Koska esiintymät ovat samassa järjestyksessä kuin Raamatun tekstissä, halutun kohdan löytämisen ei pitäisi olla vaikeaa.

Salaman järjestelmässä eri tavoin haetut sanat merkitään tulokseen erilaisin sulkumerkein riippuen siitä, minkälaista hakuavainta käytetään. Näin on mahdollista tuottaa listoja löydettyjen sanojen esiintymistiheydestä. Tätä menetelmää on käytetty yllä olevien taulukoiden laadinnassa.

Hakujärjestelmää voi tarvittaessa kehittää edelleen. Edellä jo mainittiin semanttisen koodauksen lisääminen rikastettuun tekstiin. Toinen mahdollinen kehityskohde on niin sanottujen monisanaisten ilmausten lisääminen. Raamattu ei kuitenkaan sisällä runsaasti idiomeja ja muita monisanaisia ilmauksia. Lisäksi järjestelmässä on jo nyt olemassa mahdollisuus hakea kahta tai kolmea perättäistä sanaa joko pintamuodon tai lemman perusteella. Mikäli käyttäjät kokevat tarpeelliseksi merkitä erikseen monisanaiset ilmaisut, myös tämä toiminto voidaan liittää Raamattu-hakuun.

Tarkkaa hakumenetelmää voidaan käyttää myös silloin, kun hakuja tehdään hyvin laajoihin teksteihin.[8] Tällöin haku suoritetaan kahdessa osassa siten, että ensin haetaan tekstistä ne virkkeet, joissa haettu sana esiintyy. Tulos sisältää myös haun kannalta turhaa aineistoa. Tämä supistettu teksti kuitenkin prosessoidaan analysaattorin ja disambiguaattorin avulla sellaiseen muotoon, josta tarkka haku voidaan tehdä.

Painetun Raamatun hakusanakirjan ja digitaalisen hakujärjestelmän vertailu osoittaa, että käsityönä koottu manuaalinen hakusanakirja on monella tavalla vaillinainen. Kattavaa hakujärjestelmää ei millään järkevällä tavalla voi valmistaa painotuotteeksi. Lisäksi massiivisen painotuotteen käyttäminen olisi hidasta ja hankalaa. Digitaalinen järjestelmä sitä vastoin on kattava ja tarkka. Tilan puute ei pakota rajaamaan valintoja. Hakua voi myös tehdä useilla tavoilla sen mukaan, mikä menetelmä milloinkin on tarkoituksenmukaisinta. Salaman hakujärjestelmään voi tutustua osoitteessa 77.240.23.241/tagger

Kuvattu hakujärjestelmä toimii toistaiseksi yksityisellä palvelimella eikä se ole yleisessä käytössä.

FT Arvi Hurskainen (arvi.hurskainen@helsinki.fi) on täysin palvellut Helsingin yliopiston professori. Hän on tutkinut kieliteknologiaa vuodesta 1985 lähtien ja kehittänyt tutkimuksen pohjalta erilaisia tiedonhakuun liittyviä sovelluksia.

8 Hurskainen 2021a, b.

KIRJALLISUUS

Hurskainen, Arvi (2019). Intelligent Search Engines. *Technical Reports on Language Technology.* Report No. 45. http://www.njas.helsinki.fi/salama/intelligent-search-engines.pdf.

Hurskainen, Arvi (2021a). Accurate information retrieval from large corpora: Non-extended monosyllabic Swahili verbs. *Technical reports on language technology.* Report No. 67. http://www.njas.helsinki.fi/salama/accurate-information-retrieval-from-large-corpora-1.pdf.

Hurskainen, Arvi (2021b). Accurate information retrieval from large corpora: Extended monosyllabic Swahili verbs. *Technical reports on language technology*. Report No. 68. http://www.njas.helsinki.fi/salama/accurate-information-retrieval-from-large-corpora-2.pdf.

Hurskainen, Arvi (2021c). Evaluation of four search systems of Finnish Bible. *Technical reports on language technology*. Report No. 69. http://www.njas.helsinki.fi/salama/evaluation-of-four-search-systems-of-finnish-bible.pdf.

Vuorela, Vilho (1962a). *Raamatun hakusanakirja I: Vanha Testamentti.* Porvoo: WSOY.

Vuorela, Vilho (1962b). *Raamatun hakusanakirja II: Uusi Testamentti.* Porvoo: WSOY.