Technical Reports in Language Technology
Report No 50, 2020
http://www.njas.helsinki.fi/salama

Revolutionizing information retrieval from the Finnish Bible[1]

Arvi Hurskainen
Department of World Cultures, Box 59
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

The report compares the coverage and accuracy of the hand-compiled search system of Vilho Vuorela to the Finnish Bible translation, and the computational search system developed within the framework of Salama. The Salama search system has full coverage and also full accuracy. The hand-compiled reference work has been compiled considering the assumed need of the users, and heavy selection is made. Although Vuorela's work is poor in coverage, it is accurate and it has probably managed to fulfil the basic needs of the users. However, the computational search system, that is based on the analysis and disambiguation of text, is superior in all aspects.

Key Words: *morphology, information retrieval.*

1 Introduction

Those who have studied theology, and also many laymen, have become acquainted with the two-volume reference work of Vilho Vuorela. The reference work, based on the Bible translation of 1938, has been an indispensable aid for decades for those, who want to be acquainted with the Bible. The digital technology, however, has brought new kinds of search methods. The traditional string search method may be known to most people. In this method, information is searched on the basis of surface words or parts of the words.
The inflected forms of words, however, make the information search difficult. The formation of the optimal search key is often difficult. The search result may be defective, and often there is also something, which was not intended. The good search system has two criteria, which it should fulfil. The search should be covering, that is, all searched for hits should be found. On the other hand, the search should be accurate, that is, the result should have only those hits, which were searched for.

Is it possible to achieve such search results? And if it is possible, through which methods? This report will deal with this question.

An accurate and covering search method can be achieved through the analysis and disambiguation of the text. This text form can then be modified so that each word of the original text is attached to its base form and part-of-speech code.

When the target text is enriched so that after each word there is the lemma of that word plus its part-of-speech code, we get such a text form, which makes accurate search possible.
An example of an enriched text form:

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

In this report I will show the differences of the hand made and computerized search systems. The comparison is made by using the reference work of Vuorela and the computational search approach of Salama, which includes several search systems. Because the reference work of Vuorela concerns the translation of years 1933/1938, the comparison is made using the Salama search engine adapted to the same translation. Further, because Vuorela has separate volumes for the Old Testament and New Testament, also the Salama search engine searches these two sections of the Bible separately.

Comparison was made with three kinds of material. A lemma list was first computed from the Bible. This list was then divided into proper names and ordinary words. Each of these two lists was further divided into two lists. In one list type, the lemmas were arranged according to their frequency. In another list type, lemmas were shuffled into arbitrary order. The frequency list makes it possible to search for most common words. The shuffled list makes it possible to take objective extracts from any part of the list. The third comparison method is to study such words, which are considered to be among the most commonly searched words. What we lose on objectivity in the last method, we gain in interest.

2 Comparison of two search systems

The comparison results of the search systems are displayed in table form. The frequency lists were produced with Salama, which can produce accurate lists. This list in itself contains the coverage information of Salama, and no further study is needed. The coverage of Vuorela's reference work was found by counting frequencies manually. Only such entries were counted, which had some context. Plain references without context were excluded.

2.1 Most common words in Bible

In the tables below, occurrences of the most common words in Bible are displayed. The table contains two sections. On the left, statistics produced with Salama are displayed, On the right, corresponding statistics of Vuorela are displayed. The last column shows the percentual coverage of Vuorela for each word.

Table 1 contains 20 such words in Bible, which are among the most common words.

Table 1.
SALAMA Vuorelan hakusanakirja

All	VT	UT	VT	UT	All	%
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

Table 1 shows that it is not possible to list the occurrence of common words in a printed work. Only a small fraction of occurrences is listed. However, some words are considered more important than others, which is understandable.

2.2 Randomly selected words

Next we see how randomly selected words have been described in Vuorela and how covering the descriptions are. We take 20 such words from the beginning of the shuffled list, which occur at least three times in Bible (Table 2).

Table 2.

SALAMA			Vuorelan hakusanakirja			
All	VT	UT	VT	UT	All	%
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

Rare words are better represented in Vuorela than more common words. No word is fully described.

2.3 Most common proper names

The situation with proper names is similar with ordinary words. Only a fraction of occurrences is listed (Table 3). In addition, some common names are poorly represented. Of special notice are such names as Jeesus, Saul, Aaron, Salomo and Joosua, for which only a small part of occurrences is listed.

Table 3.
SALAMA Vuorelan hakusanakirja

All	VT	UT	VT	UT	All	%
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

2.4 Randomly selected proper names

The situation with randomly selected proper names is rather grim. Most of the proper names are not listed at all in Vuorela. On the other hand, only one common proper name is in the extract (Table 4). Such words, which occur only once, were removed from the list.

Table 4.
SALAMA Vuorelan hakusanakirja

All	VT	UT	VT	UT	All	%
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

2.5 Commonly searched words

I do not know which words in Bible are among the most searched words. To this sample (Table 5) I have selected such words, which I most likely would search, assuming that also others do the same.

Table 5.

SALAMA			Vuorelan hakusanakirja			
All	VT	UT	VT	UT	All	%
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

The coverage of these words in Vuorela is much higher than of the words in the randomly selected list. It is likely that also Vuorela has considered these words important and listed many examples of them. However, only two words, *vanhurkauttaa* and *kadotus* have full coverage.

3 Evaluation of the search systems

The fundamental difference between Salama search system and the reference books of Vuorela is in their coverage. The former finds all words regardless their surface form or number. Vuorela has used varying methods when selecting the words and when deciding how many examples should be listed. Only very seldom all occurrences are listed.
Although Vuorela's method is not covering, it is rather accurate. No wrong examples are listed. Furthermore, Vuorela has subdivided some important words into subclasses, which

serves users. It would be possible to make sub-divisions of words with Salama, but it would require adding semantic tags.

Another important difference between these two search systems is related to objectivity. In manual search, various disturbing factors may distort the result. Also personal biases of the writer affect the outcome. Also varying working conditions may affect the work. In printed works, the maximum size sets often absolute limits, and the compiler is forced to make selection.

One can claim that digital search is objective, when no selection is needed. It is an entirely different thing to consider, whether such search method is always sensible. If the result contains thousands of hits, it may be tedious to find the precise information needed.

Fortunately, search can be made with many more methods than by using the lemma form as search key. Search can be targeted also to surface text, and search can be constrained in various ways. It is also possible to search for more than one word, by using such operators as AND and OR.

Search can also be made on the basis of two or three consecutive words. This can be done using surface words as key, or lemma forms as key. When the lemma forms are used as key, all such hits will be found, which have the same sequence of lemma forms, regardless their surface forms.

The digital search system has also the advantage, that search results can be copied to the user's own document. Because the hits are displayed in the order where they occur in Bible, it is easy to scroll the screen to the desired point.

In the Salama system, the words searched in different ways are marked with codes, which show the type of search used. Three kinds of parentheses are used, {}, [], and <>. As a consequence, various types of reference lists can be produced. This method was used in producing the statistics in the above tables.

The search system can be further developed in various ways. Above I mentioned the semantic codes added to the enriched text. Another possibility is the isolation of multiword expressions in the analysed text. However, the Bible does not have many idioms or other types of multiword expressions. In addition, it is already possible to search for two or three consecutive words.

The comparison of the printed reference work and a digital search system shows, that the manually compiled compendium is in many ways defective. It is not feasible to produce a covering printed reference work. Also, the use of a massive printed work would be clumsy and slow. The digital sear system is covering and precise, and free of space limitations. Search can be done in several ways, depending on search task. Salama search system is located in the address 77.240.23.241/tagger.

The search system described above is on a private server and not publicly available.

References

Hurskainen, Arvi (2019), Intelligent search engines. *Technical reports in language technology*. Report No. 45. http://www.njas.helsinki.fi/salama/intelligent-search-engines.pdf

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

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