Stemming and Lemmatization in the Clustering of Finnish Text Documents

Tuomo Korenius[1], Jorma Laurikkala[1], Kalervo Järvelin[2], Martti Juhola[1]
{ [1]Department of Computer Sciences, [2]Center for Advanced Studies },
FIN-33014 University of Tampere,
Finland
{ Tuomo Korenius, Jorma.Laurikkala, Kalervo.Jarvelin, Martti Juhola }@uta.fi

ABSTRACT
Stemming and lemmatization were compared in the clustering of Finnish text documents. Since Finnish is a highly inflectional and agglutinative language, we hypothesized that lemmatization, involving splitting of the compound words, would be more appropriate normalization approach than the straightforward stemming. The relevance of the documents were evaluated with a four-point relevance assessment scale, which was collapsed into binary one by considering all the relevant and only the highly relevant documents relevant, respectively. Experiments with four hierarchical clustering methods supported the hypothesis. The stringent relevance scale showed that lemmatization allowed the single and complete linkage methods to recover especially the highly relevant documents better than stemming. In comparison with stemming, lemmatization together with the average linkage and Ward's methods produced higher precision. We conclude that lemmatization is a better word normalization method than stemming, when Finnish text documents are clustered for information retrieval.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering.*

General Terms
Algorithms, Experimentation, Standardization

Keywords
Normalization, Stemming, Lemmatization, Clustering.

1. INTRODUCTION
In this paper, we look at *word form normalization* for the document clustering in a highly inflectional language, Finnish. We shall consider two word form normalization methods, stemming and lemmatization. Below we shall first consider earlier research on stemming and lemmatization and then the feature of the Finnish language, which motivate the evaluation of the effectiveness of stemming versus lemmatization for document clustering.

CIKM'04, November 8–13, 2004, Washington, DC, USA.

Stemming has been the most widely applied morphological technique for information retrieval. With stemming, the searcher does not need to worry about the correct truncation point of search keys. Stemming also reduces the total number of distinct index entries. Further, stemming causes query expansion by bringing word variants, derivations included, together [1-3]. Some early research results with English collections questioned the effectiveness of stemming [4]. Later results by, for example, [2] and [5], find stemming useful especially when long enough retrieved lists of documents are analyzed. Hull [5] also found that stemming is always useful with short queries. With short queries and short documents, a derivational stemmer is most useful, but with longer ones the derivational stemmer brings in more non-relevant documents. Stemming increases search key ambiguity and greedy stemming may be counter-productive: with long queries and documents, relevant material can be identified with conservative stemming. In languages other than English, stemmers have been even more successful than in English text retrieval – for example, in Slovenian [6], French [7], modern Greek [8], Arabic [9], and Swedish [10].

Lemmatization is another normalization technique: for each inflected word form in a document or request, its basic form, the lemma, is identified. The benefits of lemmatization are the same as in stemming. In addition, when basic word forms are used, the searcher may match an exact search key to an exact index key. Such accuracy is not possible with truncated, ambiguous stems. Homographic word forms cause ambiguity (and precision) problems – this may also occur with inflectional word forms [1]. Another problem is owing to words that cannot be lemmatized, because the lemmatizer's dictionary does not contain them.

Compound words form a special problem area in lemmatization. A compound word (or a compound) is a word formed from two or more component words [11]. Often no difference is made between the compounds in which the components are spelled together and the compounds in which the components are spelled separately. However, in information retrieval, this distinction is essential. Therefore, we refer by compound word to the case in which the components are spelled together.

Compounds may be split into their components in lemmatization. When indexing a text collection, both compounds and their components may be recorded in the database index thus enabling retrieval through all combinations of compound components. Based on [1] and [2] it seems beneficial to use lemmas instead of stems. This may not be the case in all languages, but seems a reasonable conclusion, considering how different Finnish and English morphologically are [1]. Alkula's [1] findings also suggest that compound splitting is beneficial for retrieval.

Although Willett [12] did not consider stemming in his review, stemming together with the clustering of English documents seems reasonable, because it may improve the retrieval performance. Since stemming may also reduce the size of the index file, it is especially beneficial for complex clustering methods. For example, El-Hamdouchi and Willett [13] utilized stemmed collections in their thorough comparison of clustering algorithms. Sometimes the use of stemming in the clustering of English text varies according to the aims of the study: Nilsson [14] opted for minimal pre-processing to ease the analysis of results.

Previous results [6-10] suggest that stemming might be also valuable in the clustering of documents in some other languages than English. However, it may not be an optimal approach to the clustering of agglutinative languages. Firstly, stemmers do not conflate compounds whenever the first components do not match exactly. Secondly, they are unable to split compounds, which typically have the head-modifier structure and the headword is the last and more important component for clustering. A recent study by Rosell [10] lends support to this conclusion: Although stemming increased both the internal and external quality of clusters, a hybrid approach combining stemming and compound splitting gave clearly better results than stemming alone, when Swedish text documents were grouped with the *k*-means technique.

We hypothesize that lemmatization with split compounds provides a better clustering result than stemming in a highly inflectional and agglutinative language like Finnish. We study the performance of four hierarchical clustering methods with stemmed and lemmatized data and also compare the clustering results to the nearest-neighbor retrieval. Moreover, we perform the evaluation in a test collection providing graded relevance assessments on a four-point scale [15-17]. Thus, we are able to analyze whether the possible performance differences are connected to document quality.

2. MATERIALS

2.1 The Finnish Language

The *Finnish language* is a very inflectional and compound rich language. Its *inflectional* and *derivational morphology* is considerably more complex than that of the Indo-European languages like English. If Finnish text words are stored in their inflected forms, this results in clearly greater space requirements for Finnish text compared to that of English texts of corresponding length. For example, Finnish has more case endings than is usual in the Indo-European languages. Finnish case endings correspond to prepositions or postpositions in other languages (cf. Finnish *auto*/*ssa*, *auto*/*sta*, *auto*/*on*, *auto*/*lla* and English in the car, out of the car, into the car, by the car). There are 15 cases, while English has only two [18].

In Finnish, several layers of endings may be affixed to word stems, indicating number, case, possession, modality, tense, person, and other morphological characteristics. This results in an enormous number of possible distinct word forms: a noun may have some 2,000 forms, an adjective 6,000, and a verb 12,000 forms. Moreover, these figures do not include the effect of derivation, which increases the figures roughly by a factor of 10 [19]. Consonant gradation makes the inflection even more complicated, as the stem of a word may alter when certain types of endings are attached to it. For example, the word *laki* (law) has in practice four inflected stems: *laki*-, *lake*-, *lai*-, and *lae*-. The common root of the stems consists of only two characters, which renders it inappropriate as a search term.

Several languages, Germanic and Finno-Ugrian languages included, are rich in compounds in contrast to English, which is phrase-oriented. For example, The Dictionary of Modern Standard Finnish contains some 200,000 entries, of which two-thirds are compound words [20, p. 68]. For example the English phrase 'Turnover Tax Bureau' is *liike*|*vaihto*|*vero*|*toimisto* in Finnish (word boundaries here marked by '|'). In Finnish, compounding results in a problem of retrieving the second or later elements of compounds, for example *verotoimisto* (tax bureau), if the searcher is not able to recall all possible first components.

2.2 The Test Collection

We applied clustering techniques to a collection of 53,893 articles published in three Finnish newspapers in 1988-1992 [15]. The articles were mainly of domestic and foreign affairs and economics. The average article length was 233 words. Most paragraphs consisted of two or three sentences.

The relevance of 16,539 documents to 30 queries had earlier been assessed with a four-level relevance scale [15-17]. These documents were judged as

- *irrelevant* - the document was not about the topic of the query (N = 14,588),
- *marginally relevant* - the topic was mentioned (N = 886),
- *fairly relevant* - the topic was discussed briefly (N = 700), and
- *highly relevant* - the topic was the main theme of the article (N = 366).

The articles were assessed by two experienced journalists and two information retrieval specialists. The relevance of documents to 20 queries was evaluated by two or three persons and the rest 10 queries by one person. The assessors agreed in 73% of the assessments. Differences of one point (21%) were solved by taking the assessment from each judge in turn, and the most plausible grade after reevaluation was selected, when differences of two or three points (6%) were settled. The remaining 37,353 documents were considered irrelevant.

2.3 Preprocessing of the Collection

Since we did not have access to a computer with a large enough central memory to process the whole collection, a sample of 5,000 documents was taken. The sample contained the relevant documents and 3,074 irrelevant ones. To compensate for the possible differences in the degree of the irrelevance, the irrelevant documents were randomly selected among the graded and non-graded ones.

The sample was processed with two Finnish morphological analysis programs. The stemmed version of the sample was created with the Snowball [21] stemmer for Finnish. Only suffixes were removed from the compounds, because, as a Porter stemmer, the program was unable to split the compound words. The dimensionality of the document-term matrix was reduced from

5,000×84,567 to 5,000×13,616 by removing terms that appeared in less than five documents. Stop words were not removed, because many stems are words of a stop word list.

FinTwol[1] [20] program was applied to lemmatize the sample. Inflected words were transformed into their morphological basic forms, lemmas, so that for each lemma was included into the database index [1, 15]. The compound words were split into their components, and the components were further processed as individual lemmas. Some words, such as typing errors, were not recognized because of the dictionary-based lemmatization. However, these words were rare and their elimination probably only marginally influenced the results. The final 5,000×13,693 document-term matrix was produced with the elimination of stop words and terms that appeared in less than five documents.

Since the matrices were quite large after the frequency-based term elimination, their dimensionality was further reduced using the *principal components analysis* (PCA) [22]. PCA is a well-known statistical technique which creates new variables (principal components) from the old ones by combining them. Although PCA is a standard approach to data reduction, it has not been used extensively in the area of information retrieval. Since the application of PCA to the lemmatized sample has been described elsewhere [23, 24], it is unnecessary to repeat here all the details.

PCA was applied to mean-centered data and the principal components which accounted for 80% of the total variance were selected in the order of magnitude. Data were not standardized because of the equitable variances. The stemmed sample was processed with the PCA method as the lemmatized sample. The samples reduced considerably: the stemmed and lemmatized samples could be described with 1,834 and 1,500 principal components instead of the respective 13,616 and 13,693 original variables.

3. METHODS

3.1 Hierarchical Clustering

We applied here the single linkage, complete linkage, group average linkage, and Ward's clustering methods [25, 26]. Although more effective clustering techniques (for example, the principal direction divisive partitioning [14]) are available, the hierarchical clustering techniques were used, because they have often been included in the cluster-based search studies [12, 13, 25]. The selected methods cluster in an agglomerative manner: The starting point are the individual objects, which are grouped together into larger clusters, until all the objects are in the same cluster. The consecutive fusions are represented with a tree structure known as the *dendrogram* [26] which gives is a visual as well as a mathematical presentation of the clustering process.

Traditionally, the dendrograms have been utilized in the cluster-based search as search trees, where the search can proceed in a top-down or bottom-up manner until a node (cluster) of sufficient quality is found [12]. Since this type search has been found only comparable to the simpler search methods [13], we chose a different approach. The tree was cut at optimal height and the resulting partition was used to facilitate the cluster-based search.

[1] Trial version available at http://www.lingsoft.fi/cgi-bin/fintwol/

Selecting the optimal number of clusters is one of the central problems in cluster analysis. Although a number of formal methods have been proposed, their accuracy varies widely [26]. We applied the inconsistency coefficient [27] to the dendrograms created by the hierarchical methods. The inconsistency coefficient is similar to the well-known Mojena upper tail rule [26]. Both methods utilize statistics of the previous fusion levels, but the inconsistency coefficient considers only some of the previous levels (typically two or three), while the Mojena rule considers them all.

3.2 Retrieval Performance Evaluation

The search capability of the cluster analysis methods was evaluated using the recall, precision, and effectiveness measures. The effectiveness measure $E=1-((1+\beta^2)PR)/(\beta^2P+R)$ [13] combines recall (R) and precision (P). The value of parameter $\beta>0$ indicates how many times more important recall is compared to precision. If $\beta=1$, recall and precision are equally important. Recall is weighed more than precision, when $\beta>1$ and vice versa. In the following, the effectiveness measure is reported with $\beta=0.5$ (precision twice as important as recall) and $\beta=2.0$ (recall twice as important as precision).

It should be noted that the effectiveness values are conversely interpreted to the recall and precision values. The *small* values indicate high effectiveness, while the large values are a sign of low effectiveness.

To calculate the performance measures, the graded relevance assessments were binarized with two different approaches. Firstly, the traditional *liberal* relevance scale was established by weighing the relevant documents against the irrelevant ones. Secondly, a *stringent* relevance scale was constructed by labeling only the highly relevant documents as relevant and all the other documents as irrelevant.

4. RESULTS

4.1 Cutting the Dendrograms

The reduced samples were clustered with the four clustering methods and the resulting dendrograms were cut at the heights indicated by the inconsistency coefficient. The optimal partitions obtained from the stemmed sample with the single linkage, complete linkage, average linkage, and Ward's method contained 1,812 (z=2), 1,357 (z=3), 300 (z=2), and 129 (z=3) clusters, respectively. All the dendrograms of the lemmatized data, except those of the complete linkage method, were clearly cut further away from the root than those of the stemmed data. Therefore, the corresponding partitions of the lemmatized sample included 3,197 (z=2), 537 (z=2), 668 (z=3), and 160 (z=3) clusters. Since the number of previous fusion levels (z) was selected experimentally, it was either two or three in the computation of the inconsistency coefficient.

The cluster size distributions were as expected. The single linkage method produced, due to the *chaining effect* [26], some large and various singleton clusters, while the complete linkage method created many compact clusters. The average linkage method produced more large clusters than the two former methods. The Ward's method discovered many equally-sized large clusters.

4.2 Optimal Clusters

To quantify the optimal performance of the cluster-based search, the cluster with the highest recall on the stringent scale was selected as the best search result for each query. Recall was used instead of precision, because clusters far too small to be considered as useful search results typically had high precision.

Table 1 shows performance measures according to recall on the stringent relevance scale. The single linkage, average linkage, and Ward's methods achieved higher recall and lower precision with the stemmed than lemmatized data. The performance of the complete linkage methods was the other way around. Lemmatization led to more effective optimal performance than stemming, because the optimal lemmatized clusters had more balanced recall and precision than the optimal stemmed clusters.

Table 1. The medians of the performance measures of the optimal clusters produced with the single linkage (SL), complete linkage (CL), average linkage (AL), and Ward's (WA) methods from the stemmed and lemmatized samples. Relevance was assessed with the *stringent* scale.

Statistics	SL	CL	AL	WA
Stemmed sample				
Recall	0.50	0.33	0.93	0.79
Precision	0.16	0.37	0.11	0.16
Effectiveness (β=2)	0.76	0.66	0.67	0.56
Lemmatized sample				
Recall	0.39	0.62	0.77	0.73
Precision	0.70	0.25	0.25	0.20
Effectiveness (β=2)	0.62	0.56	0.51	0.53

Table 2. The medians of the performance measures of the optimal clusters produced with the single linkage (SL), complete linkage (CL), average linkage (AL), and Ward's (WA) methods from the stemmed and lemmatized samples. Relevance was assessed with the *liberal* scale.

Statistics	SL	CL	AL	WA
Stemmed sample				
Recall	0.12	0.13	0.74	0.52
Precision	1.00	1.00	0.47	0.70
Effectiveness (β=2)	0.92	0.84	0.42	0.52
Lemmatized sample				
Recall	0.10	0.29	0.39	0.50
Precision	1.00	0.91	0.87	0.85
Effectiveness (β=2)	0.88	0.69	0.56	0.46

As expected, recall decreased and precision increased, when the performance of the optimal clusters was measured with the liberal relevance scale in place of the stringent one (see Table 2). In general, stemming produced again higher recall and lower precision than lemmatization. Due to the high recall, the average linkage method was more effective with the stemmed than lemmatized sample.

The optimal clusters of the average linkage and Ward's methods showed promising performance and were also of reasonable size. Furthermore, these methods grouped many of the marginally and fairly relevant documents together with the highly relevant ones. This can be seen from Table 2: The median recall of the average linkage and Ward's methods were 0.74 and 0.52 and 0.39 and 0.50 on the stemmed and lemmatized samples, respectively.

Table 2 also shows that the optimal clusters were mainly made of the relevant data. The optimal clusters created from the stemmed and lemmatized samples contained 47-100% and 85-100% relevant documents, respectively. However, the best clusters discovered with the single and complete linkage methods were often quite small.

4.3 Search performance

The practical performance of the cluster-based search was evaluated with a simple search engine which computed the cosine similarity between the query and the mean of each cluster, and returned the cluster most similar to the query to the user.

The single, complete, and average linkage methods had clear difficulties in searching the highly relevant documents from the stemmed sample (see Table 3). When the most relevant documents were retrieved, the cluster was usually optimal, but 20 and 9 of the single and complete linkage clusters contained no highly relevant documents. The optimal clusters were successfully retrieved, when also all relevant documents were considered relevant.

The optimal complete, average, and Ward's clusters were very accurately retrieved from the lemmatized sample. The majority of the retrieved clusters were optimal as measured both with the liberal and stringent relevance scales. Table 3 shows that when the clusters were ranked according to their relevance to each query, 50% of the retrieved single and complete linkage clusters and 75% of the average linkage and Ward's clusters were the best, i.e. were ranked the first. Again, many of the retrieved single linkage clusters did not contain highly relevant documents.

Table 4 gives the statistics of the sizes of clusters retrieved from the stemmed and lemmatized samples. Expectedly, the majority of the retrieved single linkage clusters were very small. The size distribution of the clusters retrieved from the stemmed sample was even more skew towards the small clusters than that of the lemmatized sample. The average linkage and Ward's methods created larger clusters from the stemmed sample than from the lemmatized one. Also, the sizes of the clusters of the stemmed sample were not as equitable as those of the lemmatized sample. The stemmed sample included a number of large clusters. Conversely to the other methods, complete linkage grouped the stemmed sample into smaller clusters than the lemmatized sample.

Table 3. The five point summary of the ranks of the retrieved optimal clusters created by the single linkage (SL), complete linkage (CL), average linkage (AL), and Ward's (WA) methods from the stemmed and lemmatized samples. Ranks both with the liberal (L) and stringent (S) relevance assessments are given. *N* = the number of retrieved clusters that contained relevant documents.

Statistics	SL		CL		AL		WA	
Relevance	L	S	L	S	L	S	L	S
Stemmed sample								
Minimum	1	1	1	1	1	1	1	1
Lower quartile	2	1	1	1	1	1	1	1
Median	3	1	2	2	1	1	1	1
Upper quartile	3	2	4	2	1	1	1	1
Maximum	4	3	7	6	4	2	7	3
N	25	10	25	21	26	23	29	28
Lemmatized sample								
Minimum	1	1	1	1	1	1	1	1
Lower quartile	1	1	1	1	1	1	1	1
Median	1	1	1	1	1	1	1	1
Upper quartile	3	2	2	2	1	1	1	1
Maximum	5	3	5	3	3	2	5	3
N	27	20	30	30	29	27	29	28

Table 4. The five point summary of the sizes of retrieved clusters created by applying the single linkage (SL), complete linkage (CL), average linkage (AL), and Ward's (WA) methods to the stemmed and lemmatized samples.

Statistics	SL	CL	AL	WA
Stemmed sample				
Minimum	1	1	1	15
Lower quartile	1.00	3.00	12.00	26.50
Median	1.00	5.00	40.00	51.50
Upper quartile	3.50	10.00	120.00	94.00
Maximum	174	26	234	622
Lemmatized sample				
Minimum	1	4	3	14
Lower quartile	1.00	11.00	16.75	20.00
Median	4.00	18.00	24.50	31.50
Upper quartile	11.25	29.25	42.50	43.00
Maximum	61	88	112	92

The performance measures for the cluster-based search are depicted in Tables 5 and 6, where the medians of the recall, precision, and effectiveness measures for the 30 queries are reported. Table 5 utilizes the liberal relevance scale, whereas the stringent relevance is used in Table 6.

Stemming and lemmatization were compared by assessing the differences in the medians. The positive differences in the recall and precision medians indicate that the results from the lemmatized data were better than those of the stemmed data. Conversely, since smaller effectiveness values are better, the negative differences of the effectiveness medians imply that the lemmatized data produced better results. For example, the difference of 0.69-0.89=-0.20 shows that the complete linkage based search was markedly more effective (β=2.0) with the lemmas (see Table 5).

The two-tailed Wilcoxon signed ranks test [28] showed that all significant (p<0.05) differences in the medians of performance measures were in favor of lemmatization. As shown in Table 5, lemmatization improved clearly the performance of the single and complete linkage methods as compared to stemming. Both methods had higher recall and effectiveness on the liberal relevance assessment scale. The average linkage and Ward's methods achieved considerably higher precision levels with the lemmatized data. Consequently, the precision weighed effectiveness improved clearly as well.

When viewed on the stringent relevance scale, results were somewhat different (see Table 6). The single linkage method showed even better general improvement in performance with lemmatized data than earlier. Also, the complete linkage method achieved higher recall, but the two normalization techniques had the same effect on precision. Average linkage gained smaller improvement in precision, but effectiveness was better with both weighs. Neither recall nor precision of the Ward's method was improved by lemmatization, when only the highly relevant documents were considered relevant.

Nearest neighbor (NN) searching [29] was used as a comparison method. The same number of documents as in the corresponding cluster was retrieved with the NN technique for each query. These search results are shown in Tables 7 and 8, where, for example, column title 'NN_{SL}' refers to the performance values obtained from the 30 NN searches, each of which included as many documents as the corresponding single linkage cluster.

On the liberal relevance assessment scale the NN searching showed the same pattern of differences between the stemmed and lemmatized samples as the cluster-based searches (see Table 7). Table 8 shows the median differences between the two normalization methods, when the NN searching was studied on the stringent relevance scale. The patter of differences was similar between the NN method and the clustering methods excluding the Ward's method. The NN technique found differences, when it searched the same number of documents as in the clusters retrieved with the Ward's method.

Table 5. The medians of performance measures for the single linkage (SL), complete linkage (CL), average linkage (AL), and Ward's (WA) clustering based searches from stemmed (S) and lemmatized (L) samples. The numbers of positive (# L>S) and negative ranks (# L<S), and *p* value are given from the Wilcoxon test. Significant ($p<0.05$) differences in the medians are shown in bold font. Relevance was assessed with the *liberal* scale.

Statistics	SL	CL	AL	WA
Recall on the liberal scale				
Median L	0.05	0.27	0.48	0.52
Median S	0.03	0.09	0.67	0.53
Difference	**0.02**	**0.18**	-0.19	-0.01
# L>S / # L<S	17/6	25/3	9/16	13/15
p	0.04	<0.001	0.19	0.90
Precision on the liberal scale				
Median L	1.00	0.90	0.87	0.83
Median S	1.00	1.00	0.44	0.57
Difference	0	-0.10	**0.43**	**0.26**
# L>S / # L<S	7/5	8/15	26/3	18/9
p	0.26	0.96	<0.001	0.01
Effectiveness (β=0.5) on the liberal scale				
Median L	0.78	0.41	0.33	0.34
Median S	0.89	0.68	0.62	0.55
Difference	**-0.11**	**-0.27**	**-0.29**	**-0.21**
# L>S / # L<S	6/18	4/24	5/24	8/21
p	0.01	<0.001	<0.001	0.01
Effectiveness (β=2.0) on the liberal scale				
Median L	0.93	0.69	0.48	0.46
Median S	0.97	0.89	0.43	0.53
Difference	**-0.04**	**-0.20**	0.05	-0.07
# L>S / # L<S	7/20	3/25	14/15	11/18
p	0.01	<0.001	0.84	0.26

Table 6. The medians of performance measures for the single linkage (SL), complete linkage (CL), average linkage (AL), and Ward's (WA) clustering based searches with stemmed (S) and lemmatized (L) samples. The numbers of positive (# L>S) and negative ranks (# L<S), and *p* value are given from the Wilcoxon test. Significant ($p<0.05$) differences in the medians are shown in bold font. Relevance was assessed with the *stringent* scale.

Statistics	SL	CL	AL	WA
Recall on the stringent scale				
Median L	0.15	0.50	0.78	0.82
Median S	0	0.13	0.90	0.84
Difference	**0.15**	**0.37**	-0.12	-0.02
# L>S / # L<S	14/3	20/2	9/11	8/9
p	0.02	<0.001	0.78	0.95
Precision on the stringent scale				
Median L	0.27	0.21	0.18	0.18
Median S	0	0.20	0.09	0.12
Difference	**0.27**	0.01	**0.09**	0.06
# L>S / # L<S	13/5	13/14	24/4	17/12
p	0.01	0.80	<0.001	0.16
Effectiveness (β=0.5) on the stringent scale				
Median L	0.77	0.78	0.79	0.79
Median S	1.00	0.83	0.89	0.86
Difference	**-0.23**	-0.05	**-0.10**	-0.07
# L>S / # L<S	2/16	11/16	4/24	12/17
p	<0.001	0.15	<0.001	0.22
Effectiveness (β=2.0) on the stringent scale				
Median L	0.82	0.61	0.50	0.55
Median S	1.00	0.86	0.71	0.73
Difference	**-0.18**	**-0.25**	**-0.21**	-0.18
# L>S / # L<S	3/15	7/20	5/23	10/19
p	<0.01	<0.001	<0.001	0.10

Table 7. The medians of performance measures for the nearest neighbor searching from the stemmed (S) and lemmatized (L) samples. The numbers of positive (# L>S) and negative ranks (# L<S), and *p* value are given from the Wilcoxon test. Significant ($p<0.05$) differences in the medians are shown in bold font. Relevance was assessed with the *liberal* scale.

Statistics	NN_{SL}	NN_{CL}	NN_{AL}	NN_{WM}
Recall on the liberal scale				
Median L	0.05	0.27	0.45	0.46
Median S	0.02	0.12	0.57	0.49
Difference	0.03	**0.15**	-0.12	-0.03
# L>S / # L<S	17/6	25/2	12/18	13/17
p	0.06	<0.001	0.13	0.51
Precision on the liberal scale				
Median L	1.00	0.88	0.81	0.82
Median S	1.00	1.00	0.54	0.60
Difference	0	-0.12	**0.27**	**0.22**
# L>S / # L<S	9/6	8/18	26/3	22/8
p	0.19	0.21	<0.001	<0.01
Effectiveness (β=0.5) on the liberal scale				
Median L	0.78	0.40	0.38	0.42
Median S	0.91	0.60	0.58	0.48
Difference	**-0.13**	**-0.20**	**-0.20**	**-0.06**
# L>S / # L<S	6/19	4/24	3/27	8/22
p	0.01	<0.001	<0.001	<0.01
Effectiveness (β=2.0) on the liberal scale				
Median L	0.93	0.69	0.51	0.51
Median S	0.98	0.86	0.45	0.51
Difference	**-0.05**	**-0.17**	0.06	0
# L>S / # L<S	7/17	3/25	17/13	14/16
p	0.04	<0.001	0.82	0.48

Table 8. The medians of performance measures for the nearest neighbor searching from the stemmed (S) and lemmatized (L) samples. The numbers of positive (# L>S) and negative ranks (# L<S), and *p* value are given from the Wilcoxon test. Significant ($p<0.05$) differences in the medians are shown in bold font. Relevance was assessed with the *stringent* scale.

Statistics	NN_{SL}	NN_{CL}	NN_{AL}	NN_{WM}
Recall on the stringent scale				
Median L	0.17	0.50	0.67	0.62
Median S	0.01	0.16	0.76	0.64
Difference	**0.16**	**0.34**	-0.09	-0.02
# L>S / # L<S	15/6	25/1	11/11	10/8
p	0.01	<0.001	0.53	0.59
Precision on the stringent scale				
Median L	0.26	0.27	0.17	0.17
Median S	0.03	0.26	0.10	0.12
Difference	0.23	0.01	**0.07**	**0.05**
# L>S / # L<S	16/9	14/14	25/5	20/8
p	0.27	0.64	<0.001	0.01
Effectiveness (β=0.5) on the stringent scale				
Median L	0.77	0.69	0.79	0.80
Median S	0.96	0.77	0.88	0.86
Difference	**-0.19**	**-0.08**	**-0.09**	**-0.06**
# L>S / # L<S	6/19	12/16	5/25	8/20
p	0.01	0.04	<0.001	0.03
Effectiveness (β=2.0) on the stringent scale				
Median L	0.80	0.60	0.55	0.62
Median S	0.98	0.84	0.70	0.72
Difference	**-0.18**	**-0.24**	**-0.15**	**-0.10**
# L>S / # L<S	6/18	5/23	6/24	8/20
p	<0.01	<0.001	<0.001	0.02

5. DISCUSSION

A sample of 5,000 documents from a Finnish newspaper article collection was clustered with four agglomerative hierarchical methods to experimentally evaluate whether lemmatization produces better results than stemming. The cluster-based search results were compared with the searching enabled with the nearest neighbor (NN) technique. The collection was assessed into marginally, fairly, and highly relevant or irrelevant documents with respect to 30 queries. Therefore, the performance measures could be evaluated with the traditional liberal relevance scale and the stringent relevance scale, where only the highly relevant documents were considered as relevant.

The optimal clustering results lend some support to our hypothesis that of the two word normalization techniques lemmatization would produce better results in the clustering of Finnish text. Generally, the optimal stemmed clusters had higher recall and lower precision than the optimal lemmatized clusters both on the liberal and stringent relevance scale. Although recall was weighed in the effectiveness measure, lemmatization led on the whole to more effective search than stemming, because recall and precision associated with lemmatization were better balanced.

The average linkage and Ward's methods produced quite good optimal results. These methods typically found from the stemmed sample 93% and 79%, and from lemmatized sample 77% and 73%, of the highly relevant documents (see Table 1). Moreover, many of the fairly and marginally relevant documents were grouped together with the highly relevant ones: The average linkage and Ward's methods were usually able to retrieve 74% and 52% of the relevant documents in the stemmed sample, while the 39% and 50% of the relevant objects could be recovered from the lemmatized sample (see Table 2). Furthermore, the median precision values of Table 2 indicate that the optimal clusters were mostly made of the relevant documents.

The search results are clearly consistent with our assumption of the better performance of lemmatization in the clustering of Finnish text. The optimal clusters of the lemmatized sample were successfully retrieved with the simple search engine. For half the queries, the optimal single and complete linkage were discovered, and, moreover, 75% of the optimal average linkage and Ward's clusters were returned (see Table 3). Retrieval of the optimal stemmed clusters was difficult: The single and complete linkage based searches missed often clusters containing highly relevant documents.

All the significant differences in the medians of the recall, precision, and effectiveness measures favored lemmatization (see Tables 5-8). When the results were studied on the liberal relevance scale, the clustering and NN enabled searching showed similar differences between the two word normalization methods (see Tables 5 and 7). The worst performers, the single and complete linkage methods, were more effective with lemmas, because searches enabled with them found more relevant documents from the lemmatized sample than stemmed one. The average linkage and Ward's methods capitalized the benefits of lemmatization as the considerably higher precision levels.

Lemmatization seemed to help the single and complete linkage methods to find especially the highly relevant documents (see Table 6). The average linkage method achieved again better precision from the lemmatized data, but Ward's method did not show any significant differences between the word normalization methods, when only the highly relevant documents were defined relevant. The NN searching discovered generally the same pattern of significant differences as clustering, excluding the Ward's method (see Table 8).

Besides the search results, the size distributions of the retrieved clusters suggest that the clustering methods were not able to group the stemmed sample as well as the lemmatized one (see Table 4). It is obvious that the single linkage clustering had problems with the both types of normalization, but since half of the clusters retrieved from the stemmed sample contained only one document, stemming was clearly a weak choice. Also, the typical complete linkage clusters were too small to be considered suitable search results. The average linkage and Ward's method showed the ability to recover reasonable-sized clusters both from the stemmed and lemmatized sample. However, the size distributions of the retrieved lemmatized clusters were clearly more balanced than those of the retrieved stemmed clusters.

The poor performance of the single and complete linkage methods with stems was as anticipated, because stemming was likely to enhance the typical features of these methods to their extremes. Single linkage groups first together the similar objects, while the highly different objects are merged to the large cluster in the final stages of the clustering process [13, 26]. The complete linkage method defines the between-cluster distance in a stricter manner than single linkage. Therefore, complete linkage method creates typically small tightly-bound clusters [13, 26].

Stemming increases the similarity between several documents, because as the different word forms are brought together, the number of discriminating variables decreases. On the other hand, a number of documents may become much different from the other documents because of the stemmer's inability to process the compound words, which are regular in the Finnish text. Consequently, single linkage builds an extremely skew dendrogram, cutting of which produces an abundance of small clusters accompanied with some very large clusters. Complete linkage produces a more balanced dendrogram, but the highly different objects are likely to produce unusually small clusters.

Our results are consistent with studies by Alkula [1] and Krovetz [2] suggest that the lemmatization may be a better approach than stemming. In addition, the results agree with Alkula's [1] and Rosell's [29] results indicating that compound splitting is beneficial for retrieval of texts written in agglutinative languages.

6. CONCLUSIONS

Two word normalization methods, stemming and lemmatization, were compared in the clustering of text documents in Finnish, which is a highly inflectional and agglutinative language. On the basis of the experimental results obtained with four hierarchical clustering methods, we conclude that lemmatization is better in conjunction with the clustering of Finnish text than stemming.

Lemmas were found better both on the liberal and strict relevance scales, which were established by collapsing the four-point relevance assessment scale. Lemmatization helped the single and complete linkage methods to recover especially the highly relevant documents, whereas the average linkage and Ward's methods achieved increased precision. The major drawback of

stemming was probably the inability to split compound words that are spelled together in Finnish.

Future research could investigate the clustering of the non-normalized Finnish text. It is obvious that, compared to lemmatization, the results would be worse, but it might be interesting to study the differences in the performance of searches from the stemmed and original documents.

7. ACKNOWLEDGMENTS

The first and second authors are grateful for the Academy of Finland for financial support (grants 50973 and 55243).

8. REFERENCES

[1] Alkula, R. From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software. *Information Retrieval*, *4*, (2001), 195-208.

[2] Krovetz, R. Viewing morphology as an inference process. In *Proceedings of the Sixteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)* (Pittsburg, PA, 27 June - 1 July 1993). ACM Press, New York, NY, 1993, 191-202.

[3] Pirkola, A. Morphological typology of languages for information retrieval. *Journal of Documentation*, *57*, 3 (2001), 330-348.

[4] Harman D. How effective is suffixing? *Journal of the American Society for Information Science*, *42*, 1 (1991), 7-15.

[5] Hull, D. Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*, *47*, 1 (1996), 70-84.

[6] Popovic, M., and Willett, P. The effectiveness of stemming for natural-language access to Slovene textual data. *Journal of the American Society for Information Science*, *43*, 1 (1992), 384-390.

[7] Savoy, J. A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science*, *50*, 10 (1999), 944-952.

[8] Kalamboukis, T. Z. Suffix stripping with modern Greek. *Program*, *29*, 3 (1995), 313-321.

[9] Abu-Salem, H., Al-Omari, M., and Evens, M. W. Stemming methodologies over individual query words for an Arabic information retrieval system. *Journal of the American Society for Information Science*, *50*, 6 (1999), 524-529.

[10] Rosell, M. Improving clustering of Swedish newspaper articles using stemming and compound splitting. In *Fourteenth Nordic Conference on Computational Linguistics (NoDaLiDa 2003)* (Reykjavik, Island, May 30-31, 2003). http://www.nada.kth.se/~rosell/publications/papers/improvingClustering03.pdf

[11] Matthews, P. H. *The Concise Oxford Dictionary of Linguistics*. Oxford University Press, Oxford - New York, NY, 1997.

[12] Willett, P. Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management*, *24*, 5 (1988), 577-597.

[13] El-Hamdouchi, A., and Willett, P. Comparison of hierachic agglomerative clustering methods for document retrieval. *The Computer Journal*, *32*, 3 (1989), 220-227.

[14] Nilsson, M. Hierarchical clustering using non-greedy principal direction divisive partitioning. *Information Retrieval*, *5*, 4 (2002), 311-321.

[15] Kekäläinen, J. *The Effects of Query Complexity, Expansion and Structure on Retrieval Performance in Probabilistic Text Retrieval*. Ph.D. Thesis, University of Tampere, 1999. Acta Universitatis Tamperensis, vol. 678.

[16] Sormunen, E. *A Method for Measuring Wide Range Performance of Boolean Queries in Full-Text Databases*. Ph.D. Thesis, University of Tampere, 2000. Acta Universitatis Tamperensis, vol. 748.

[17] Kekäläinen, J., and Järvelin, K. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, *53*, 13 (2002), 1120-1129.

[18] Karlsson, F. *Finnish grammar*. WSOY, Porvoo, 1987.

[19] Koskenniemi, K. An application of the two-level model to Finnish. In *Computational morphosyntax: Report on research 1981-84*. Publications 13, University of Helsinki, Department of General Linguistics, Helsinki, 1985, 19-41.

[20] Koskenniemi, K. *Two-level morphology: A general computational model for word-form recognition and production*. Publications 11, University of Helsinki, Department of General Linguistics, Helsinki, 1983.

[21] Porter, M. F. *Snowball: A language for stemming algorithms*, 2001. http://snowball.tartarus.org/

[22] Jolliffe, I. T. *Principal Components Analysis*. Springer-Verlag, New York, 1986.

[23] Korenius, T., Laurikkala, J., and Juhola, M. On applying the principal components analysis and cosine similarity for information retrieval. A manuscript submitted to *Information Processing & Management*.

[24] Korenius, T., Laurikkala, J., Juhola, M., and Järvelin, K. Hierarchical clustering of a Finnish newspaper article collection with graded relevance assessments. A manuscript submitted to *Information Retrieval*.

[25] Rasmussen, E. Clustering algorithms. In *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, Upper Saddle River, NJ, 1992, 419-442.

[26] Everitt, B. S., Landau, S., and Leese, M. *Cluster Analysis*. Arnold, London, 2001.

[27] The Math Works Inc. *Statistics Toolbox User's Guide*. The Math Works Inc., Natick, 2002.

[28] Pett, M. A. *Nonparametric Statistics for Health Care Research*. Sage Publications, Thousand Oaks, CA, 1997.

[29] Mitchell, T. M. *Machine Learning*. McGraw-Hill, New York, 1997.