Biomedical Text Retrieval in Languages with a Complex Morphology

Stefan Schulz [a] **Martin Honeck** [a] **Udo Hahn** [b]

[a] Department of Medical Informatics, Freiburg University Hospital
http://www.imbi.uni-freiburg.de/medinf

[b] Text Knowledge Engineering Lab, Freiburg University
http://www.coling.uni-freiburg.de

Abstract

Document retrieval in languages with a rich and complex morphology – particularly in terms of derivation and (single-word) composition – suffers from serious performance degradation with the stemming-only query-term-to-text-word matching paradigm. We propose an alternative approach in which morphologically complex word forms are segmented into relevant subwords (such as stems, named entities, acronyms), and subwords constitute the basic unit for indexing and retrieval. We evaluate our approach on a large biomedical document collection.

1 Introduction

Morphological alterations of a search term have a negative impact on the recall performance of an information retrieval (IR) system (Choueka, 1990; Jäppinen and Niemistö, 1988; Kraaij and Pohlmann, 1996), since they preclude a *direct* match between the search term proper and its morphological variants in the documents to be retrieved. In order to cope with such variation, morphological analysis is concerned with the reverse processing of inflection (e.g., *'search⊕ed'*, *'search⊕ing'*)[1], derivation (e.g., *'search⊕er'* or *'search⊕able'*) and composition (e.g., German *'Blut⊕hoch⊕druck'* [*'high blood pressure'*]). The goal is to map all occurring morphological variants to some canonical base form — e.g., *'search'* in the examples from above.

The efforts required for performing morphological analysis vary from language to language. For English, known for its limited number of inflection patterns, lexicon-free general-purpose stemmers (Lovins, 1968; Porter, 1980) demonstrably improve retrieval performance. This has been reported for other languages, too, dependent on the generality of the chosen approach (Jäppinen and Niemistö, 1988; Choueka, 1990; Popovic and Willett, 1992; Ekmekçioglu et al., 1995; Hedlund et al., 2001; Pirkola, 2001). When it comes to a broader scope of morphological analysis, including derivation and composition, even for the English language only restricted, domain-specific algorithms exist. This is particularly true for the medical domain. From an IR view, a lot of specialized research has already been carried out for medical applications, with emphasis on the lexico-semantic aspects of dederivation and decomposition (Pacak et al., 1980; Norton and Pacak, 1983; Wolff, 1984; Wingert, 1985; Dujols et al., 1991; Baud et al., 1998).

While one may argue that single-word compounds are quite rare in English (which is not the case in the medical domain either), this is certainly not true for German and other basically agglutinative languages known for excessive single-word nominal compounding. This problem becomes even more pressing for technical sublanguages, such as medical German (e.g., *'Blut⊕druck⊕mess⊕gerät'* translates to *'device for measuring blood pressure'*). The problem one faces from an IR point of view is that besides fairly standardized nominal compounds, which already form a regular part of the sublanguage proper, a myriad of *ad hoc* compounds are formed on the fly which cannot be anticipated when formulating a retrieval query though they appear in relevant documents. Hence, enumerating morphological variants in a semi-automatically generated lexicon, such as proposed for French (Zweigenbaum et al., 2001), turns out to be infeasible, at least for German and related languages.

[1] '⊕' denotes the string concatenation operator.

Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain, Philadelphia, July 2002, pp. 61-68. Association for Computational Linguistics.

Furthermore, medical terminology is characterized by a typical mix of Latin and Greek roots with the corresponding host language (e.g., German), often referred to as *neo-classical compounding* (McCray et al., 1988). While this is simply irrelevant for general-purpose morphological analyzers, dealing with such phenomena is crucial for any attempt to cope adequately with medical free-texts in an IR setting (Wolff, 1984).

We here propose an approach to document retrieval which is based on the idea of segmenting query and document terms into basic subword units. Hence, this approach combines procedures for deflection, dederivation and decomposition. Subwords cannot be equated with linguistically significant morphemes, in general, since their granularity may be coarser than that of morphemes (cf. our discussion in Section 2). We validate our claims in Section 4 on a substantial biomedical document collection (cf. Section 3).

2 Morphological Analysis for Medical IR

Morphological analysis for IR has requirements which differ from those for NLP proper. Accordingly, the decomposition units vary, too. Within a canonical NLP framework, linguistically significant *morphemes* are chosen as nondecomposable entities and defined as the smallest content-bearing (*stem*) or grammatically relevant units (*affixes* such as prefixes, infixes and suffixes). As an IR alternative, we here propose *subwords* (and grammatical affixes) as the smallest units of morphological analysis. Subwords differ from morphemes only, if the meaning of a combination of linguistically significant morphemes is (almost) equal to that of another nondecomposable medical synonym. In this way, subwords preserve a sublanguage-specific composite meaning that would get lost, if they were split up into their constituent morpheme parts.

Hence, we trade linguistic atomicity against medical plausibility and claim that the latter is beneficial for boosting the system's retrieval performance. For instance, a medically justified minimal segmentation of *'diaphysis'* into *'diaphys⊕is'* will be preferred over a linguistically motivated one (*'dia⊕phys⊕is'*), because the first can be mapped to the quasi-synonym stem *'shaft'*. Such a mapping would not be possible with the overly unspecific morphemes *'dia'* and *'phys'*, which occur in numerous other contexts as well (e.g.*'dia⊕gnos⊕is'*, *'phys⊕io⊕logy'*). Hence, a decrease of the precision of the retrieval system would be highly likely due to over-segmentation of semantically opaque compounds. Accordingly, we distinguish the following decomposition classes:

Subwords like {*'gastr'*, *'hepat'*, *'nier'*, *'leuk'*, *'diaphys'*, ...} are the primary content carriers in a word. They can be prefixed, linked by infixes, and suffixed. As a particularity of the German medical language, proper names may appear as part of complex nouns (e.g., *'Parkinson⊕verdacht'* [*'suspicion of Parkinson's disease'*]) and are therefore included in this category.

Short words, with four characters or less, like {*'ion'*, *'gene'*, *'ovum'*}, are classified separately applying stricter grammatical rules (e.g., they cannot be composed at all). Their stems (e.g., *'gen'* or *'ov'*) are *not* included in the dictionary in order to prevent false ambiguities. The price one has to pay for this decision is the inclusion of derived and composed forms in the subword dictionary (e.g., *'anion'*,*'genet'*,*'ovul'*).

Acronyms such as {*'AIDS'*, *'ECG'*, ...} and *abbreviations* (e.g., *'chron.'* [for *'chronical'*], *'diabet.'* [for *'diabetical'*]) are nondecomposable entities in morphological terms and do not undergo any further morphological variation, e.g., by suffixing.

Prefixes like {*'a-'*, *'de-'*, *'in-'*, *'ent-'*, *'ver-'*, *'anti-'*, ...} precede a subword.

Infixes (e.g., *'-o-'* in *"gastr⊕o⊕intestinal"*, or *'-s-'* in *'Sektion⊕s⊕bericht'* [*'autopsy report'*]) are used as a (phonologically motivated) 'glue' between morphemes, typically as a link between subwords.

Derivational suffixes such as {*'-io-'*, *'-ion-'*, *'-ie-'*, *'-ung-'*, *'-itis-'*, *'-tomie-'*, ...} usually follow a subword.

Inflectional suffixes like {*'-e'*, *'-en'*, *'-s'*, *'-idis'*, *'-ae'*, *'-oris'*, ...} appear at the very end of a composite word form following the subwords or derivational suffixes.

Prior to segmentation a language-specific orthographic normalization step is performed. It maps German umlauts *'ä'*, *'ö'*, and *'ü'* to *'ae'*, *'oe'*, and *'ue'*, respectively, translates *'ca'* to *'ka'*, etc. The morphological segmentation procedure for German

in January 2002 incorporates a *subword dictionary* composed of 4,648 subwords, 344 proper names, and an *affix list* composed of 117 prefixes, 8 infixes and 120 (derivational and inflectional) suffixes, making up 5,237 entries in total. 186 stop words are not used for segmentation. In terms of domain coverage the subword dictionary is adapted to the terminology of clinical medicine, including scientific terms, clinicians' jargon and popular expressions. The subword dictionary is still in an experimental stage and needs on-going maintenance. Subword entries that are considered strict synonyms are assigned a shared identifier. This thesaurus-style extension is particularly directed at foreign-language (mostly Greek or Latin) translates of source language terms, e.g., German *'nier'* EQ Latin *'ren'* (EQ English *'kidney'*), as well as at stem variants.

The morphological analyzer implements a simple word model using regular expressions and processes input strings following the principle of 'longest match' (both from the left and from the right). It performs backtracking whenever recognition remains incomplete. If a complete recognition cannot be achieved, the incomplete segmentation results, nevertheless, are considered for indexing. In case the recognition procedure yields alternative complete segmentations for an input word, they are ranked according to preference criteria, such as the minimal number of stems per word, minimal number of consecutive affixes, and relative semantic weight.[2]

3 Experimental Setting

As document collection for our experiments we chose the CD-ROM edition of MSD, a German-language handbook of clinical medicine (MSD, 1993). It contains 5,517 handbook-style articles (about 2.4 million text tokens) on a broad range of clinical topics using biomedical terminology.

In our retrieval experiments we tried to cover a wide range of topics from clinical medicine. Due to the importance of searching health-related contents both for medical professionals and the general public we collected two sets of user queries, viz. expert queries and layman queries.

[2]A semantic weight w=2 is assigned to all subwords and some semantically important suffixes, such as *'-tomie'* ['*-tomy'*] or '*-itis'*; w=1 is assigned to prefixes and derivational suffixes; w=0 holds for inflectional suffixes and infixes.

Expert Queries. A large collection of multiple choice questions from the nationally standardized year 5 examination questionnaire for medical students in Germany constituted the basis of this query set. Out of a total of 580 questions we selected 210 ones explicitly addressing clinical issues (in conformance with the range of topics covered by MSD). We then asked 63 students (between the 3rd and 5th study year) from our university's Medical School during regular classroom hours to formulate free-form natural language queries in order to retrieve documents that would help in answering these questions, assuming an ideal search engine. Acronyms and abbreviations were allowed, but the length of each query was restricted to a maximum of ten terms. Each student was assigned ten topics at random, so we ended up with 630 queries from which 25 were randomly chosen for further consideration (the set contained no duplicate queries).

Layman Queries. The operators of a German-language medical search engine (`http://www.dr-antonius.de/`) provided us with a set of 38,600 logged queries. A random sample (n=400) was classified by a medical expert whether they contained medical jargon or the wording of laymen. Only those queries which were univocally classified as layman queries (through the use of non-technical terminology) ended up in a subset of 125 queries from which 27 were randomly chosen for our study.

The judgments for identifying relevant documents in the whole test collection (5,517 documents) for each of the 25 expert and 27 layman queries were carried out by three medical experts (none of them was involved in the system development). Given such a time-consuming task, we investigated only a small number of user queries in our experiments. This also elucidates why we did not address inter-rater reliability. The queries and the relevance judgments were hidden from the developers of the subword dictionary.

For unbiased evaluation of our approach, we used a home-grown search engine (implemented in the PYTHON script language). It crawls text/HTML files, produces an inverted file index, and assigns salience weights to terms and documents based on a simple *tf-idf* metric. The retrieval process relies on the vector space model (Salton, 1989), with the cosine measure expressing the similarity between a

query and a document. The search engine produces a ranked output of documents.

We also incorporate proximity data, since this information becomes particularly important in the segmentation of complex word forms. So a distinction must be made between a document containing *'append⊕ectomy'* and *'thyroid⊕itis'* and another one containing *'append⊕ic⊕itis'* and *'thyroid⊕ ectomy'*. Our proximity criterion assigns a higher ranking to adjacent and a lower one to distant search terms. This is achieved by an *adjacency offset*, o_a, which is added to the cosine measure of each document. For a query Q consisting of n terms, $Q = t_1, t_2, ..., t_n$, the minimal distance between a pair of terms in a document, (t_i, t_j), is referred to by $d_{t_{i,j}}$. The offset is then calculated as follows:

$$o_a = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{i-1} \frac{1}{|d_{t_{i,j}}|} \quad (1)$$

We distinguished four different conditions for the retrieval experiments, *viz.* plain token match, trigram match, plain subword match, and subword match incorporating synonym expansion:

Plain Token Match (WS). A direct match between text tokens in a document and those in a query is tried. No normalizing term processing (stemming, etc.) is done prior to indexing or evaluating the query. The search was run on an index covering the entire document collection (182,306 index terms). This scenario serves as the baseline for determining the benefits of our approach.[3]

Trigram Match (TG). As an alternative lexicon-free indexing approach (which is more robust relative to misspellings and suffix variations) we considered each document and each query indexed by all of their substrings with character length '3'.

Subword Match (SU). We created an index building upon the principles of the subword approach as described in Section 2. Morphological segmentation yielded a shrunk index, with 39,315 index terms remaining. This equals a reduction rate of 78% compared with the number of text types in the collection.[4]

[3]This is a reasonable baseline, since up to now there is no general-purpose, broad-coverage morphological analyzer for German available, which forms part of a standard retrieval engine.

[4]The data for the English version, 50,934 text types with 24,539 index entries remaining after segmentation, indicates a significantly lower reduction rate of 52%. The size of the English subword dictionary (only 300 entries less than the German one) does not explain the data. Rather this finding reveals that the English corpus has fewer single-word compounds than the German one.

Synonym-Enhanced Subword Match (SY). Instead of subwords, synonym class identifiers which stand for several subwords are used as index terms.

The following add-ons were supplied for further parametrizing the retrieval process:

Orthographic Normalization (O). In a preprocessing step, orthographic normalization rules (cf. Section 2) were applied to queries and documents.

Adjacency Boost (A). Information about the position of each index term in the document (see above) is made available for the search process.

Table 1 summarizes the different test scenarios.

Name of Test	Index Made of	Orthographic Normalization	Adjacency Boost
WS	Words	-	-
WSA	Words	-	+
WSO	Words	+	-
WSAO	Words	+	+
TG	Trigrams	-	-
SU	Subwords	+	+
SY	Synonym Class Ids	+	+

Table 1: Different Test Scenarios

4 Experimental Results

The assessment of the experimental results is based on the aggregation of all 52 selected queries on the one hand, and on a separate analysis of expert vs. layman queries, on the other hand. In particular, we calculated the average interpolated precision values at fixed recall levels (we chose a continuous increment of 10%) based on the consideration of the top 200 documents retrieved. Additionally, we provide the average of the precision values at all eleven fixed recall levels (11pt recall), and the average of the precision values at the recall levels of 20%, 50%, and 80% (3pt recall).

We here discuss the results from the analysis of the complete query set the data of which is given in Table 2 and visualized in Figure 1. For our baseline (*WS*), the direct match between query terms and document terms, precision is already poor at low recall points ($R \leq 30\%$), ranging in an interval from 53.3% to 31.9%. At high recall points ($R \geq 70\%$),

	Precision (%)						
Rec. (%)	WS	WSA	WSO	WS AO	TG	SU	SY
0	53.3	56.1	53.3	60.0	54.8	**74.0**	**73.2**
10	46.6	50.7	46.1	55.8	45.4	62.3	61.0
20	37.4	40.1	37.0	42.1	32.1	52.3	51.7
30	31.9	33.2	31.5	34.5	26.3	**45.8**	**45.1**
40	28.9	30.4	28.0	30.3	20.2	39.2	36.5
50	26.6	28.6	26.0	28.7	15.9	**35.6**	32.7
60	24.5	25.9	23.5	25.0	9.3	29.7	28.1
70	19.1	19.9	17.9	18.7	6.5	24.4	22.7
80	14.4	15.2	13.0	14.0	4.4	**19.6**	18.1
90	9.5	9.8	9.6	9.9	0.8	**14.7**	**14.6**
100	3.7	3.9	3.8	4.0	0.64	**10.0**	**10.2**
3pt avr	26.1	28.0	25.3	28.3	17.4	35.8	34.1
11pt avr	26.9	28.5	26.3	29.4	19.6	37.0	35.8

Table 2: Precision/Recall Table for All Queries

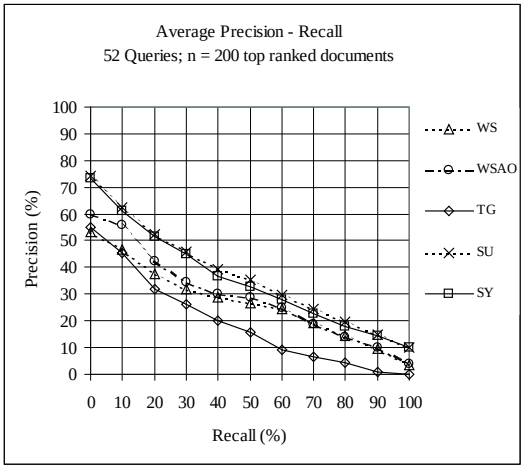

Figure 1: Precision/Recall Graph for All Queries

precision drops from 19.1% to 3.7%. When we take term proximity (adjacency) into account (*WSA*), we observe a small though statistically insignificant increase in precision at all recall points, 1.6% on average. Orthographic normalization only (*WSO*), however, caused, interestingly, a marginal decrease of precision, 0.6% on average. When both parameters, orthographic normalization and adjacency, are combined (*WSAO*), they produce an increase of precision at nine from eleven recall points, 2.5% on average compared with *WS*. None of these differences are statistically significant when the two-tailed Wilcoxon test is applied at all eleven recall levels.

Trigram indexing (*TG*) yields the poorest results of all methodologies being tested. It is comparable to *WS* at low recall levels ($R \leq 30\%$), but at high ones its precision decreases almost dramatically. Unless very high rates of misspellings are to be expected (this explains the favorable results for trigram indexing in (Franz et al., 2000)) one cannot really recommend this method.

The subword approach (*SU*) clearly outperforms the previously discussed approaches. We compare it here with *WSAO*, the best-performing lexicon-free method. Within this setting, the gain in precision for *SU* ranges from 6.5% to 14% ($R \leq 30\%$), while for high recall values ($R \geq 70\%$) it is still in the range of 4.8% to 6%. Indexing by synonym class identifiers (*SY*) results in a marginal decrease of overall performance compared with *SU*. To estimate the statistical significance of the differences *SU* vs. *WSAO* and *SY* vs. *WSAO*, we compared value pairs at each fixed recall level, using the two-tailed Wilcoxon test (for a description and its applicability for the interpretation of precision/recall graphs, cf. (Rijsbergen, 1979)). Statistically significant results ($\alpha < 5\%$) are in bold face in Table 2.

The data for the comparison between expert and layman queries is given in the Tables 3 and 4, respectively, and they are visualized in the Figures 2 and 3, respectively. The prima facie observation that layman recall data is higher than those of the experts is of little value, since the queries were acquired in quite different ways (cf. Section 3). The adjacency criterion for word index search (*WSA*) has no influence on the layman queries, probably because they contain fewer search terms. This may also explain the poor performance of trigram search. A considerably higher gain for the subword indexing approach (*SU*) is evident from the data for layman queries. Compared with *WSAO*, the average gain in precision amounts to 9.6% for layman queries, but only 5.6% for expert queries. The difference is also obvious when we compare the statistically significant differences ($\alpha < 5\%$) in both tables (bold face). This is also compatible with the finding that the rate of query result mismatches (cases where a query did not yield any document as an answer) equals zero for *SU*, but amounts to 8% and 29.6% for expert and laymen queries, respectively, running under the token match paradigm *WS** (cf. Table 5).

When we compare the results for synonym class indexing (*SY*), we note a small, though statistically insignificant improvement for layman queries at some recall points. We attribute the different re-

	Precision (%)						
Rec. (%)	WS	WSA	WSO	WS AO	TG	SU	SY
0	50.5	56.8	50.3	60.8	56.6	67.3	64.7
10	45.8	53.2	44.6	59.8	48.7	60.3	60.3
20	39.3	44.7	38.1	48.6	35.8	50.8	50.3
30	32.2	34.8	31.0	37.3	30.6	46.5	45.7
40	26.3	29.3	24.3	29.0	21.6	37.3	32.0
50	22.3	26.5	20.9	26.5	19.7	34.2	28.3
60	19.2	22.0	16.9	20.1	10.9	24.7	20.3
70	11.8	13.5	9.3	11.1	7.7	19.9	15.7
80	9.9	11.6	7.1	9.1	6.5	14.2	10.3
90	3.7	4.4	4.1	4.7	1.7	**9.2**	8.3
100	3.6	4.0	4.0	4.4	1.3	**8.3**	7.6
3pt avr	23.8	27.6	22.1	28.1	20.7	33.1	29.7
11pt avr	24.1	27.3	22.8	28.3	21.9	33.9	31.2

Table 3: Precision/Recall Table for Expert Queries

	Precision (%)						
Rec. (%)	WS	WSA	WSO	WS AO	TG	SU	SY
0	55.8	55.4	56.1	59.1	53.2	**80.3**	**81.0**
10	47.3	48.5	47.6	52.2	42.2	64.0	61.6
20	35.6	35.8	35.9	36.2	28.6	**53.6**	**52.9**
30	31.7	31.7	31.9	31.9	22.2	**45.1**	**44.5**
40	31.3	31.3	31.4	31.4	19.0	**41.0**	**40.7**
50	30.6	30.6	30.7	30.7	12.3	36.8	36.8
60	29.5	29.5	29.6	29.6	7.8	34.4	35.3
70	25.8	25.8	25.8	25.8	5.3	28.5	29.2
80	18.5	18.5	18.5	18.5	2.5	24.6	25.3
90	14.8	14.8	14.8	14.8	0.0	19.7	**20.5**
100	3.7	3.7	3.7	3.7	0.0	**11.5**	**12.7**
3pt avr	28.2	28.3	28.4	28.5	14.4	38.3	38.4
11pt avr	29.5	29.6	29.6	30.4	17.5	40.0	40.0

Table 4: Precision/Recall Table for Layman Queries

sults partly to the lower baseline for layman queries, partly to the probably more accentuated vocabulary mismatch between layman queries and documents using expert terminology. However, this difference is below the level we expected. In forthcoming releases of the subword dictionary in which coverage, stop word lists and synonym classes will be augmented, we hope to demonstrate the added value of the subword approach more convincingly.

Generalizing the interpretation of our data in the light of these findings, we recognize a substantial increase of retrieval performance when query and text tokens are segmented according to the principles of the subword model. The gain is still not overwhelming.

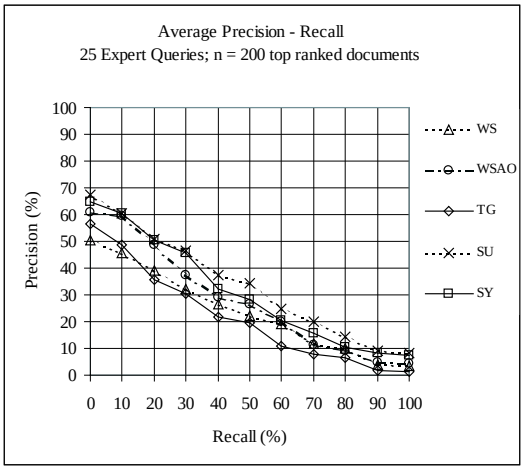

Figure 2: Precision/Recall Graph for Expert Queries

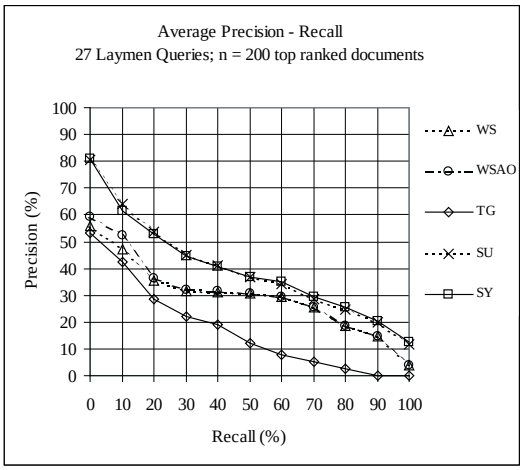

Figure 3: Precision/Recall Graph for Layman Queries

With regard to *orthographic normalization*, we expected a higher performance benefit because of the well-known spelling problems for German medical terms of Latin or Greek origin (such as in *'Zäkum'*, *'Cäkum'*, *'Zaekum'*, *'Caekum'*, *'Zaecum'*, *'Caecum'*). For our experiments, however, we used quite a homogeneous document collection following the spelling standards of medical publishers. The same standards apparently applied to the original multiple choice questions, by which the acquisition of expert queries was guided (cf. Section 3). In the layman queries, there were only few Latin or Greek terms, and, therefore, they did not take advantage of the spelling normalization. However, the experience with medical text retrieval (especially on medical reports which exhibit a high rate of spelling variations) shows that orthographic normalization is a desider-

Rate of Query / Document Mismatch (%)							
	WS	WSA	WSO	WSAO	TG	SU	SY
Exp.	8.0	8.0	8.0	8.0	0.0	0.0	0.0
Lay.	29.6	29.6	29.6	29.6	0.0	0.0	0.0
All	19.2	19.2	19.2	19.2	0.0	0.0	0.0

Table 5: Query / Document Mismatch

atum for enhanced retrieval quality. The proximity (*adjacency*) of search terms as a crucial parameter for output ranking proved useful, so we use it as default for subword and synonym class indexing.

Whereas the usefulness of *Subword Indexing* became evident, we could not provide sufficient evidence for *Synonym Class Indexing*, so far. However, synonym mapping is still incomplete in the current state of our subword dictionary. A question we have to deal with in the future is an alternative way to evaluate the comparative value of synonym class indexing. We have reason to believe that precision cannot be taken as the sole measure for the advantages of a query expansion in cases where the *subword approach* is already superior (for all layman and expert queries this method retrieved relevant documents, whereas word-based methods failed in 29.6% of the layman queries and 8% of the expert queries, cf. Figure 5). It would be interesting to evaluate the retrieval effectiveness (in terms of precision and recall) of different versions of the *synonym class indexing* approach in those cases where retrieval using word or subword indexes fails due to a complete mismatch between query and documents. This will become even more interesting when mappings of our synonym identifiers to a large medical thesaurus (MeSH, (NLM, 2001)) are incorporated into our system. Alternatively, we may think of user-centered comparative studies (Hersh et al., 1995).

4.1 The *AltaVista*^{TM} Experiment

Before we developed our own search engine, we used the *AltaVista*^{TM} *Search Engine 3.0* (`http://solutions.altavista.com`) as our testbed, a widely distributed, easy to install off-the-shelf IR system. For the conditions *WSA*, *SU*, and *SY*, we give the comparative results in Table 6. The experiments were run on an earlier version of the dictionary – hence, the different results. *AltaVista*^{TM} yielded a superior performance for all three major test scenarios compared with our home-grown engine. This is not at all surprising given all the tuning

	Precision (%)					
	AltaVista			Experimental		
Recall (%)	WSA	SU	SY	WSA	SU	SY
0	53.6	69.4	66.9	56.8	67.3	64.2
10	51.7	65.5	60.5	53.2	60.3	58.8
20	45.4	61.4	54.9	44.7	50.7	48.3
30	34.9	55.4	51.6	34.8	45.7	39.4
40	29.5	51.4	46.7	29.3	34.6	32.9
50	27.8	49.7	44.1	26.5	31.2	29.4
60	26.2	40.7	39.2	22.0	22.2	20.1
70	18.1	32.6	31.7	13.5	18.9	16.5
80	15.2	26.3	22.4	11.6	13.4	12.1
90	5.6	20.1	11.4	4.4	7.9	8.3
100	5.4	16.3	11.0	4.0	7.0	7.5
3pt avrg	29.5	45.8	40.5	27.6	32.8	29.9
11pt avrg	28.5	44.4	40.0	27.3	32.6	30.7

Table 6: Precision/Recall Table for Expert Queries comparing the *AltaVista*^{TM} with our Experimental Search Engine

efforts that went into *AltaVista*^{TM}. The data reveals clearly that commercially available search engines comply with our indexing approach. In an experimental setting, however, their use is hardly justifiable because their internal design remains hidden and, therefore, cannot be modified under experimental conditions.

The benefit of the subword indexing method is apparently higher for the commercial IR system. For *AltaVista*^{TM} the average precision gain was 15.9% for *SU* and 11.5% for *SY*, whereas our simple *tfidf*-driven search engine gained only 5.3% for *SU* and 3.4% for *SY*. Given the imbalanced benefit for both systems (other things being equal), it seems highly likely that the parameters feeding *AltaVista*^{TM} profit even more from the subword approach than our simple prototype system.

5 Conclusions

There has been some controversy, at least for simple stemmers (Lovins, 1968; Porter, 1980), about the effectiveness of morphological analysis for document retrieval (Harman, 1991; Krovetz, 1993; Hull, 1996). The key for quality improvement seems to be rooted mainly in the presence or absence of some form of dictionary. Empirical evidence has been brought forward that inflectional and/or derivational stemmers augmented by dictionaries indeed perform substantially better than those without access

to such lexical repositories (Krovetz, 1993; Kraaij and Pohlmann, 1996; Tzoukermann et al., 1997).

This result is particularly valid for natural languages with a rich morphology — both in terms of derivation and (single-word) composition. Document retrieval in these languages suffers from serious performance degradation with the stemming-only query-term-to-text-word matching paradigm.

We proposed here a dictionary-based approach in which morphologically complex word forms, no matter whether they appear in queries or in documents, are segmented into relevant subwords and these subwords are subsequently submitted to the matching procedure. This way, the impact of word form alterations can be eliminated from the retrieval procedure.

We evaluated our hypothesis on a large biomedical document collection. Our experiments lent (partially statistically significant) support to the subword hypothesis. The gain of subword indexing was slightly more accentuated with layman queries, probably due to a higher vocabulary mismatch.

References

R. Baud, C. Lovis, A.-M. Rassinoux, and J.-R. Scherrer. 1998. Morpho-semantic parsing of medical expressions. In *Proc. of the 1998 AMIA Fall Symposium*, pages 760–764.

Y. Choueka. 1990. RESPONSA: An operational full-text retrieval system with linguistic components for large corpora. In A. Zampolli, L. Cignoni, and E. C. Peters, editors, *Computational Lexicology and Lexicography. Special Issue Dedicated to Bernard Quemada. Vol. 1*, pages 181–217. Pisa: Giardini Editori E. Stampatori.

P. Dujols, P. Aubas, C. Baylon, and F. Grémy. 1991. Morphosemantic analysis and translation of medical compound terms. *Methods of Information in Medicine*, 30(1):30–35.

F. Ekmekçioglu, M. Lynch, and P. Willett. 1995. Development and evaluation of conflation techniques for the implementation of a document retrieval system for Turkish text databases. *New Review of Document and Text Management*, 1(1):131–146.

P. Franz, A. Zaiss, S. Schulz, U. Hahn, and R. Klar. 2000. Automated coding of diagnoses: Three methods compared. In *Proc. of 2000 AMIA Fall Symposium*, pages 250–254.

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

T. Hedlund, A. Pirkola, and K. Järvelin. 2001. Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language retrieval. *Information Processing & Management*, 37(1):147–161.

W. Hersh, D. Elliot, D. Hickam, S. Wolf, A. Molnar, and C. Leichtenstien. 1995. Towards new measures of information retrieval evaluation. In *Proc. of the 18th International ACM SIGIR Conference*, pages 164–170.

D. A. Hull. 1996. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1):70–84.

H. Jäppinen and J. Niemistö. 1988. Inflections and compounds: Some linguistic problems for automatic indexing. In *Proc. of the RIAO 88 Conference*, volume 1, pages 333–342.

W. Kraaij and R. Pohlmann. 1996. Viewing stemming as recall enhancement. In *Proc. of the 19th International ACM SIGIR Conference*, pages 40–48.

R. Krovetz. 1993. Viewing morphology as an inference process. In *Proceedings of the 16th International ACM SIGIR Conference*, pages 191–203.

J. Lovins. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1/2):22–31.

A. McCray, A. Browne, and D. Moore. 1988. The semantic structure of neo-classical compounds. In *SCAMC'88 – Proc. of the 12th Annual Symposium on Computer Applications in Medical Care*, pages 165–168.

MSD. 1993. *– Manual der Diagnostik und Therapie [CD-ROM]*. München: Urban & Schwarzenberg, 5th edition.

NLM. 2001. *Medical Subject Headings*. Bethesda, MD: National Library of Medicine.

L. Norton and M. Pacak. 1983. Morphosemantic analysis of compound word forms denoting surgical procedures. *Methods of Information in Medicine*, 22(1):29–36.

M. Pacak, L. Norton, and G. Dunham. 1980. Morphosemantic analysis of *-itis* forms in medical language. *Methods of Information in Medicine*, 19(2):99–105.

A. Pirkola. 2001. Morphological typology of languages for IR. *Journal of Documentation*, 57(3):330–348.

M. Popovic and P. Willett. 1992. The effectiveness of stemming for natural language access to Slovene textual data. *Journal of the American Society for Information Science*, 43(5):384–390.

M. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

C. J. van Rijsbergen. 1979. *Information Retrieval*. London: Butterworths, 2nd edition.

Gerard Salton. 1989. *Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer*. Reading, MA: Addison-Wesley.

E. Tzoukermann, J. Klavans, and C. Jacquemin. 1997. Effective use of natural language processing techniques for automatic conflation of multi-word terms: The role of derivational morphology, part of speech tagging, and shallow parsing. In *Proc. of the 20th International ACM SIGIR Conference*, pages 148–155.

F. Wingert. 1985. Morphologic analysis of compound words. *Methods of Information in Medicine*, 24(3):155–162.

S. Wolff. 1984. The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding. *Methods of Information in Medicine*, 23(4):195–203.

P. Zweigenbaum, S. Darmoni, and N. Grabar. 2001. The contribution of morphological knowledge to French MESH mapping for information retrieval. In *Proc. of the 2001 AMIA Fall Symposium*, pages 796–800.