Approaches to High Accuracy Retrieval:
Phrase-Based Search Experiments in the HARD track

Olga Vechtomova
Department of Management Sciences
University of Waterloo
Waterloo, Canada
ovechtom@engmail.uwaterloo.ca

Murat Karamuftuoglu
Department of Computer Engineering
Bilkent University
Ankara, Turkey
hmk@cs.bilkent.edu.tr

1. Introduction

Our main research focus this year was on the use of phrases (or multi-word units) in query expansion. Multi-word units (MWUs) comprise a number of lexical units, such as named entities ("United Nations"), nominal compounds ("amusement park") and phrasal verbs ('check in'). Although MWUs can belong to different parts of speech, our focus was on nominal MWUs. We used a combined syntactico-statistical approach for selecting nominal MWUs. In the first selection pass we obtained valid noun phrases, and in the second pass we used statistical measures to select strongly bound MWUs. We have experimented with using two statistical measures of selecting MWUs from text: the C-value (Frantzi and Ananiadou 1996, Vintar 2004) and the Log-Likelihood ratio (Dunning 1993). Selected MWUs were then suggested to the user for interactive query expansion. Two main research questions were studied in these experiments:

- Whether nominal MWUs which exhibit strong degree of stability in the corpus are better candidates for interactive query expansion than nominal MWUs selected by the frequency parameters of the individual terms they contain;
- Whether nominal MWUs are better candidates for interactive query expansion than single terms.

In more detail these experiments are presented in section 2.2.

The second focus of this work is on studying the effectiveness of noun phrases in document ranking. We have developed a new method of phrase-based document re-ranking, which addresses the problem of weighting overlapping phrases in documents, which in statistical IR models, such as probabilistic leads to the over-inflation of the document score. The method is described in detail in section 2.3.1. In section 3 we present the evaluation results, and in section 4 we discuss the differences in query expansion and retrieval performance between queries formulated by users with low and high familiarity with the topic.

2. System description

2.1 Baseline run

We submitted two baseline runs: in the first one (UWATbaseTD) we used all non-stopword terms, extracted from the title and description fields of the topic, in the second (UWATbaseT) – all terms from the title field only. For both runs we used Okapi BM25 search function (Sparck-Jones et al. 2000).

2.2 Clarification forms

According to track instructions, a clarification form (CL form) could be used by participants to elicit any feedback or additional search criteria from users. Our main interest in using clarification forms was to evaluate different techniques for selecting MWUs and phrases for interactive query expansion.

We used 25 top-ranked documents retrieved in the UWATbaseTD run for selecting query expansion units. In each of these documents 2 best sentences were selected using the same technique that we used in our HARD track participation last year (Vechtomova et al. 2004). Two main factors influenced sentence selection: (1) the *idf* weights of the original query terms present in the sentence, and (2) information value of the sentence, i.e. the combined *tf.idf* value of its words. We did not experiment with other passage-selection techniques or with full-text.

As last year, we applied Brill's rule-based tagger (Brill 1995) and BaseNP noun phrase chunker (Ramshaw and Marcus 1995) to extract noun phrases from these sentences. Multi-word units are then selected from the list of obtained noun phrases using different statistical techniques, described below.

Clarification form 1

MWUs are characterised foremost by relative stability in the corpus. Some of the noun phrases output by the NP chunker are chance word groupings, and not stable MWUs. We were interested in exploring the value of MWUs compared to all noun-phrases in representing useful query expansion concepts to the user. The method of selecting stable MWUs from noun phrases using C-value is outlined below.

Noun phrases output by the NP chunker are ranked by the average *idf* of their constituent terms. For each phrase we generate the list of all phrases that it subsumes, i.e. contiguous or non-contiguous combinations of words in forward order, including the original complete phrase. For each subphrase, the C-value is calculated. The C-value is a measure of stability of an n-gram in the corpus (Frantzi and Ananiadou 1996). The C-value formula we used is as follows:

$$C-value(a) = (length(a)-1)(freq(a)-\frac{t(a)}{c(a)}) \qquad (1)$$

Where:

t(a) is the frequency of the n-gram in longer n-grams;
c(a) is the number of longer n-grams including a;

All subphrases for a given phrase are ranked by the C-value. The top ranked subphrase is then used to replace the original phrase in the list of candidate query expansion terms. The original complete phrase may get a higher C-value than any of its subphrases, in which case it is kept without a change.

For example, in our experiment, the bigram "World Cup" received the highest C-value out of all its subphrases generated from the phrase "grueling IAU 100-kilometer World Cup" and as a consequence was selected for the phrase list. The obtained phrases were then ranked by their C-value, top 75 of which were shown to the user in the clarification form[1].

Clarification form 2
One of the research questions that we wanted to explore was whether phrases are better candidates for interactive query expansion than single terms. A phrase carries more context and information, so it should be easier for the user to select more good phrases than single terms for query expansion. To test this hypothesis we took the phrases we selected for the previous set of clarification forms, and produced a list of single terms by splitting them and removing duplicates. The terms were then shown to the user for selection.

Clarification form 3
In this form we included top 75 phrases which were output by the NP chunker and ranked by the average *idf* of their constituent terms. By comparing query expansion with the phrases selected from this clarification form with the phrases selected from the 1st and the 4th clarification forms we aim to answer the question whether the application of the measures of phrase stability in the corpus lead to better phrases for query expansion.

Clarification form 4
In the final set of clarification forms we experimented with selecting noun phrases using the Log-Likelihood measure (Dunning 1993). Log-Likelihood has been used extensively for the identification of statistically significant word collocations in text and has shown good results in English (e.g., Manning and Schütze 1999).

We calculate Log-Likelihood for bigrams, using the Ngram Statistics Package (Banerjee and Pedersen 2003). The phrase weighting was then done as follows: first, from each phrase output by the NP chunker we derived contiguous bigrams. For each bigram, its Log-Likelihood score was calculated. The highest Log-Likelihood score of any bigram

[1] This was the maximum number of phrases that could be displayed in the clarification form.

derived from the phrase was taken as the phrase weight. We then displayed the top 75 phrases in the CL form. This is a rather crude selection method, but it does reward phrases which contain a strongly bound collocation.

2.3 Experimental runs

Five experimental runs were conducted. Runs UWATexp1, UWATexp2, UWATexp3 and UWATexp4 used the feedback provided by the users to the 1st, 2nd, 3rd and 4th sets of clarification forms accordingly. In each run the query was constructed by splitting the phrases selected by the user from the corresponding clarification form into single terms and adding them to the original query terms. Each term in the expanded query was weighted in Okapi. The BM25 document retrieval function was used for topics requesting documents and BM250 passage retrieval function was used for topics requesting passages.

Run 5 (UWATexp5) was conducted using phrase search. Here for each topic we take the top 1000 documents retrieved in the UWATexp1 run and re-rank it using the phrase search method presented in section 2.3.1 below.

2.3.1 Phrase search

Following the interactive query expansion stage where the users select query expansion phrases, the next step is to use them in search. Intuitively using them as phrases in search should lead to better precision than if we split them into single words. One problem associated with the use of phrases in a statistical IR model, such as probabilistic (Sparck-Jones et al. 2000) is that some terms may occur in multiple phrases, for example let us assume there are two phrases in the query: "*air traffic*" and "*traffic control*", and two documents: the first containing one phrase "*air traffic control*", and the second – two phrases "*air traffic*" and "*traffic control*". How should they be weighted? If we calculate weights of each phrase in the document separately and then add them up to get the document score, as is currently done in the probabilistic model for single terms, then two documents will get equal scores. That obviously shouldn't be the case. But then how should the phrase weight be calculated for the first document? The situation gets more complex if we allow for non-contiguous word combinations, i.e. matching the following: "1 *air* 2 *traffic* 10 *control*" (numbers here denote positions of the words in text). Allowing match on non-contiguous word combinations is good for recall as it relaxes search constraints, but the distance between the phrase elements should be inversely related to the phrase weight. Therefore, the two main issues to be addressed by the phrase search algorithm are:

- remove the problem of overlapping phrases;
- reflect the distance between the phrase elements in the phrase weight.

We have developed the following phrase search algorithm, which attempts to address these problems:

In the first step we retrieve documents by using a query which consists of all single terms extracted from the user-selected phrases from the CL form 1 (UWATExp1 run).

The next step is to re-rank these documents by using phrase information. We take top 1000 documents per topic in the retrieved set, stem the terms in each document and create a document representation, consisting only of the stemmed occurrences of terms from the query in their original order and their sequential position number in text.

For each query phrase, all possible subphrases (i.e. contiguous and non-contiguous combinations of words) are generated and ranked in the descending order of their length. For each subphrase in the list we use *cgrep* – a pattern matching program for extracting minimal matching strings (Clarke 1995) to extract the minimal spans of text in the document containing the subphrase. Each time cgrep returns matching strings, they are removed from the document representation and the procedure is repeated with the same phrase. If no matching string is found, the program attempts to match the next phrase in the list, and so on. In this way we can match progressively longer spans containing the phrase or its subphrases. An example of extracted windows for the phrase "practical implementation" is given in figure 1 (the number preceded by the '#' sign is the sequential position of the following word in the original document text).

```
# 106 implementation # 120 practical
# 120 practical # 186 implementation
# 4 implementation
# 21 implementation
# 43 implementation
# 59 implementation
```

Figure 1. An example of windows extracted from a document for the phrase “practical implementation”

Windows extracted using the above method might overlap**.** Our approach is to eliminate overlaps in windows in a two-step process: (1) rank the windows by their weight and (2) remove overlapping words from the lower ranked windows.

<u>*Window weighting*</u>

In this approach the window weight is calculated from the combination of *idf* weights of individual query terms occurring in it. The following formula was used:

$$WindowWeight(a) = \sum_{i=1}^{n} idf_i \times \frac{n}{(span+1)^p} \qquad (2)$$

Where:

- i – word in the window a;
- n – number of words in the window a;
- $span = pos(n) - pos(1)$
  - where: $pos(i)$ – position number of the i^{th} word in the window;
- p – tuning parameter[2].

So, the more informative the words in the window are, the shorter the span is, and the more words there are in the window, the higher is the weight of the window.

<u>*Removing duplicate windows*</u>

After the windows are ranked, we remove overlapping words by doing pairwise comparison of all windows. If two windows have overlapping query word(s), these words are removed from the lower ranked window. The windows shown in figure 1 after the removal of overlapping words are illustrated in figure 2.

```
# 106 implementation # 120 practical
# 4 implementation
# 21 implementation
# 43 implementation
# 59 implementation
# 186 implementation
```

Figure 2. An example of windows after the removal of overlapping words.

All windows extracted from the document for every query phrase are then added to the same list, weighted using the formula (2) above and have the overlapping words removed. For each window we also keep the index of the phrase which was used to extract it.

<u>*Calculating document scores*</u>

The next step is to calculate document scores. First, for each phrase in the query we calculate its weight in the document as follows:

$$PhraseWeight(a) = \frac{(k+1)\times \sum_{w=1}^{n} WindowWeight(w)}{k \times NF + n} \qquad (3)$$

[2] Experiments showed that 0.2 gives the best performance on the HARD 2003 corpus.

Where:

w – window, extracted for the query phrase *a*;
n – number of windows extracted for the phrase *a*;
NF – document length normalisation factor (see equation 5 below).
k – phrase frequency normalisation factor[3].

The document length normalisation factor was calculated in the same way as in the BM25 document ranking function (Sparck-Jones 2000):

$$NF = (1-b) + b \times \frac{Doclen}{AveDocLen} \qquad (4)$$

Where:

Doclen – document length (word count);
AveDoclen – average document length in the corpus;
b – tuning constant[4].

Document score is then calculated as the sum of *PhraseWeight* values for all query phrases that occur in the document:

$$DocumentScore(d) = \sum_{a=1}^{n} PhraseWeight(a) \qquad (5)$$

Where: *a* – the query phrase occurring in the document *d*;
n – number of query phrases occurring in the document *d*.

Finally, the top 1000 documents in the originally retrieved set are re-ranked by the new document scores.

3. Results

The results of the document-level evaluation based on 46 topics[5] are presented in table 2. All expanded runs significantly improved the average precision (Soft-rel) over the baseline run UWATbaseTD, and all runs except UWATexp5 significantly improved P@10 (Soft-rel) over the baseline (t-test at .05 significance level).

Run	Soft-rel			Hard-rel		
	P@10	**R-precision**	**AveP**	**P@10**	**R-precision**	**AveP**
Title terms (**UWATbaseT**)	0.3089	0.2499	0.2196	0.2422	0.2298	0.2185
Baseline, Title + Description (**UWATbaseTD**)	0.42	0.3011	0.2693	0.3444	0.2744	0.2636
Single-term search, Query expansion with phrases from clarification form 1 (**UWATExp1**)	0.4889	0.3381	0.3176	0.4044	0.2971	0.2817
ExpRun1 reranked using the phrase-search algorithm (**UWATExp5**)	0.4422	0.3258	0.3233	0.3711	0.2854	0.2888
Single-term search, Query expansion with terms from clarification form 2 (**UWATExp2**)	0.48	0.3283	0.3026	0.4	0.2807	0.2695
Single-term search, Query expansion with phrases from clarification form 3 (**UWATExp3**)	0.4911	0.3352	0.3191	0.3978	0.3131	0.3128
Single-term search, Query expansion with phrases from clarification form 4 (**UWATExp4**)	0.4689	0.3256	0.3019	0.4	0.2875	0.2689

Table 2. Document-level results of the runs, averaged over all topics.

Retrieval performance of the expanded queries created from the user feedback to clarification forms 1 and 2 is very similar. This suggests that users tend to select similarly good terms, whether they are shown to them in the context of phrases or on their own. On average users selected 21 phrases from the 1st clarification form and 27 single terms

[3] Experiments showed that k=1.2 gives the best performance on the HARD 2003 corpus.
[4] Sparck-Jones et al. have experimentally determined that 0.75 gives best results on TREC data.
[5] Four topics had no relevant documents and were, therefore, excluded from the evaluation.

from the 2nd form. There were 675 phrase-terms selected only from the 1st form, 384 terms selected only from the 2nd form and 921 terms selected from both forms.

There is a very small difference between the performance of the queries from phrases selected using the average *idf* of their terms (UWATExp3) and queries from phrases selected using the measures of phrase stability in the corpus: the C-value (UWATExp1) and the Log-Likelihood ratio (UWATExp4). UWATExp3 gives somewhat better R-Precision and AveP results in Hard-rel evaluation. The R-precision (Hard-rel) of UWATExp3 is 5% higher than UWATExp1 and 9% higher than UWATexp4, neither of which is statistically significant. Similar performances of these three runs suggest that the statistical component of phrase selection does not play an important role in choosing query expansion phrases when it is combined with syntactical phrase selection techniques, such as POS-tagging and NP-chunking.

The phrase search algorithm (UWATExp5) did not demonstrate improvement over the performance of the single-term search method (UWATExp1). While average precision (Soft-rel) increased slightly, the precision at 10 documents and R-precision deteriorated. The use of phrases improved average precision in 17 topics and degraded precision in 28 topics (Soft-rel evaluation). The average gain was 56%, while the average loss was 24%. We have also analysed performance of the phrase-search and single-term search methods in topics formulated by users with little and much familiarity, which is discussed in the next section.

One of the problems that might have caused the overall low performance of the phrase-search method is that we did not set the span limit. The rationale for that was to capture not only phrasal, but also topical relations between terms. However, this approach may be more useful with long multi-topic documents, rather than short documents. Since HARD track collection consisted mainly of short news articles, this aspect of the phrase search method is unlikely to help distinguish between relevant and non-relevant documents more than single-term document retrieval techniques.

Table 3 shows the results of the passage-level evaluation, averaged over 25 topics, which requested passages as retrieval elements. UWATExp3 gave best results in all passage evaluation measures but one. The phrase search run UWATExp5 did better than the single-term search method UWATExp1 in CharRPrec by 7.8%, but similar or slightly worse in other measures.

Run	**Bpref@XChars**			**Prec@XChars**			**CharRPrec**
	6000	**12000**	**24000**	**6000**	**12000**	**24000**	
UWATBaseT	0.209	0.204	0.171	0.223	0.221	0.185	0.163
UWATBaseTD	0.207	0.198	0.189	0.232	0.23	0.228	0.166
UWATExp1	0.286	0.26	0.22	0.308	0.289	0.245	0.166
UWATExp5	0.267	0.245	0.223	0.279	0.275	0.25	0.179
UWATExp2	0.267	0.26	0.231	0.291	0.286	0.261	0.169
UWATExp3	0.285	0.272	0.243	0.304	0.306	0.268	0.192
UWATExp4	0.22	0.222	0.21	0.25	0.247	0.246	0.179

Table 3. Passage-level results

4. The effect of familiarity on phrase selection and retrieval performance

The familiarity metadata was used in the HARD track to indicate the extent to which the searchers formulating the topic were familiar with it. Out of 46 topics, which were used in the evaluation, there were 25 topics with user familiarity “little” and 21 topics with familiarity “much”.

We have analysed the effect of the searcher familiarity with the topic on two variables:

- the number of phrases selected for query expansion;
- the performance of different search methods.

We hypothesise that the more familiar the searcher is with the subject of the query, the more phrases they are able to choose for query expansion. Our experimental results support this hypothesis. In all four clarification forms users familiar with the topic selected substantially more QE terms and phrases than the less familiar users (Table 4). The difference observed in all clarification forms but one, CL1 (C-value selected phrases), was statistically significant (using t-test at .05 significance level).

Clarification form	Average number of selected phrases/terms		Difference
	Familiarity "little"	Familiarity "much"	
CL1: C-value selected phrases	19.6	24.9	27%
CL2: Single terms from CL1	24	36	49%
CL3: Ave. IDF selected phrases	15	25	67.5%
CL4: Log-likelihood selected phrases	19.7	29.6	50.3%

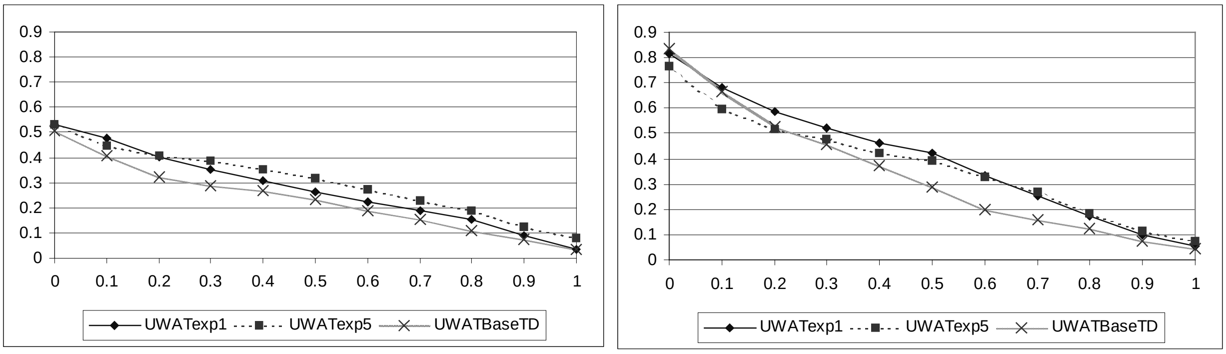
Table 4. The average number of QE phrases/terms selected by users with "little" and "much" familiarity.

Next, we hypothesise that the more familiar the searchers are with the topic**,** the better the performance of their unexpanded and expanded queries should be. The results of all baseline and experimental runs support this hypothesis: in all runs topics with "much" familiarity show higher Mean Average Precision, as shown in Table 5.

Run	Mean Average Precision (soft-rel)		Difference
	Familiarity "little"	Familiarity "much"	
Baseline, Title terms (**UWATbaseT**)	0.184	0.265	44.2%
Baseline, Title + Description (**UWATbaseTD**)	0.228	0.320	40.7%
Single-term search, Query expansion with phrases from clarification form 1 (**UWATExp1**)	0.266	0.382	43.8%
ExpRun1 reranked using the phrase-search algorithm (**UWATExp5**)	0.292	0.362	23.6%
Single-term search, Query expansion with terms from clarification form 2 (**UWATExp2**)	0.269	0.345	28.4%
Single-term search, Query expansion with phrases from clarification form 3 (**UWATExp3**)	0.280	0.368	31.2%
Single-term search, Query expansion with phrases from clarification form 4 (**UWATExp4**)	0.251	0.366	45.8%

Table 5. Mean average precision (soft-rel) of topics formulated by users with "little" and "much" familiarity.

The analysis of search results by familiarity reveals very interesting patterns in the performance of the phrase-based document re-ranking method. As mentioned in the previous section, overall the phrase-based run (UWATExp5) did not improve performance over the single-term search (UWATExp1). By analysing topics with different familiarity levels, we can see, however, that our phrase-based document re-ranking method improves the Average Precision of topics with "little" familiarity by 10%, and deteriorates the Average Precision of topics with "much" familiarity by 5.7%.



(a) **(b)**

Figure 3. Recall-precision graphs (soft-rel) of the runs for topics with (a) familiarity "little" and (b) familiarity "much".

The recall-precision graph in figure 3a shows another interesting phenomenon: for topics with familiarity "little" our phrase-based document re-ranking method has lower precision than single-term search at low recall levels and higher precision at higher recall levels, beginning from around 20% recall. Similar, but weaker pattern is evident in topics with familiarity "much" (figure 3b): phrase-based re-ranking has lower precision than single-term search up to around 60% recall level, and then starts getting slightly better. The pattern of lower precision at high recall and higher precision at high recall levels was also observed by Mitra et al. (1997) in their experiments with phrase search.

5. Conclusions

In this paper we presented a comparative evaluation of different phrase selection techniques in interactive query expansion and a phrase-based document reranking method. A combined syntactico-statistical method was used for the selection of phrases. First, noun phrases were selected using a part-of-speech tagger and a noun-phrase chunker, and secondly, different statistical measures were applied to select phrases for query expansion. Three selection methods were used: C-value, Log-Likelihood ratio and the average *idf* of phrase terms to select phrases, which were then shown to the user for interactive query expansion. Evaluation experiments did not demonstrate substantial difference between these statistical methods in their effect on the retrieval performance.

We also studied whether users select better terms when they are shown in the context of phrases, than separately. The users were asked to select query expansion items from two clarification forms: one with the complete phrases selected by the C-value, and the other with the single terms from these phrases. The two query expansion runs gave very similar results, which suggests that presenting terms in the context of phrases does not provide much help to the users in selecting good query expansion terms. However, a large proportion of terms was only selected from one of the clarification forms.

The phrase-based document re-ranking method did not demonstrate overall improvement over the single-term search technique. However, it improved the Average Precision of topics formulated by users with low familiarity. As discussed earlier in the paper, phrases differ by their stability in the corpus, therefore they should not be treated uniformly in search. For example, a document which has a partial match on a non-compositional or idiomatic phrase (e.g. "Salt Lake City") is more likely to be non-relevant, than a document that has a partial match on a non-idiomatic expression (e.g. "organic product"). Therefore the weight of the partially matching phrase should be reduced more in the first case than in the second. The continuation of this work will be to use measures of phrase stability to estimate phrase weights in the documents.

References

Banerjee, S. and Pedersen, T. (2003). The Design, Implementation and Use of the Ngram Statistics Package. Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, 2003, February, Mexico City.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging. *Computational Linguistics*, 21(4), 543-565.

Clarke, C.L.A. and Cormack, G.V. On the use of Regular Expressions for Searching Text. University of Waterloo Computer Science Department Technical Report number CS-95-07, University of Waterloo, Canada, February 1995.

Dunning, T. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), pp. 61-74, 1993.

Frantzi, K.T. and Ananiadou, S. (1996) Extracting nested collocations. In Proc. 16th Conference on Computational Linguistics, COLING, pp.41-46.

Manning C.D., Schütze H. *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts, 1999.

Mitra, M., Buckley, C., Singhal, A. and Cardie, C. An Analysis of Statistical And Syntactic Phrases. In Proceedings of RIAO97, Computer-Assisted Information Searching on the Internet, Montreal, Canada (1997) 200–214

Ramshaw, L. and Marcus, M. (1995). Text Chunking Using Transformation-Based Learning. *Proceedings of the Third ACL Workshop on Very Large Corpora*, MIT.

Sparck Jones, K., Walker, S. and Robertson, S.E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management,* 36(6), 779-808 (Part 1); 809-840 (Part 2).

Vechtomova, O., Karamuftuoglu, M., Lam, E. (2004). Interactive Search Refinement Techniques for HARD Tasks. *Proceedings of the Twelfth Text Retrieval Conference*, Voorhees, E. and Buckland, L., Editors, November 18-21, 2003, NIST, Gaithersburg, MD, pp. 820-827.

Vintar Š. (2004) Comparative Evaluation of C-Value in the Treatment of Nested Terms. In Proceedings of MEMURA 2004 Workshop (Methodologies and Evaluation of Multiword Units in Real-world Applications), Language Resources and Evaluation Conference (LREC), Lisbon, Portugal (2004) 54-57.