



# UNIVERSITY OF TAMPERE

This document has been downloaded from  
Tampub – The Institutional Repository of University of Tampere

Authors: Lehtokangas Raija, Järvelin Kalervo  
Name of article: Consistency of textual expression in newspaper articles: an argument for semantically based query expansion  
Year of publication: 2001  
Name of journal: Journal of Documentation  
Volume: 57  
Number of issue: 4  
Pages: 535-548  
ISSN: 0022-0418  
Discipline: Natural sciences / Computer and information sciences  
Language: en  
School/Other Unit: School of Information Sciences

URN: <http://urn.fi/urn:nbn:uta-3-746>

DOI: <http://dx.doi.org/10.1108/EUM0000000007104>

All material supplied via TamPub is protected by copyright and other intellectual property rights, and duplication or sale of all part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorized user.

# Consistency of Textual Expression in Newspaper Articles: An argument for semantically based query expansion

Raija Lehtokangas & Kalervo Järvelin

Department of Information Studies

FIN-33014 University of Tampere, Finland

{raija.lehtokangas, likaja}@uta.fi

## ***Abstract***

This article investigates how consistent different newspapers are in their choice of words when writing about the same news events. News articles on the same news events were taken from three Finnish newspapers and compared in regard to their central concepts and words representing the concepts in the news texts. Consistency figures were calculated for each set of three articles (the total number of sets was 60). Inconsistency in words and concepts was found between news articles from different newspapers. The mean value of consistency calculated on the basis of words was 65%; this however depended on the article length. For short news wires consistency was 83% while for long articles it was only 47%. At the concept level, consistency was considerably higher, ranging from 92% to 97% between short and long articles. The articles also represented three categories of topic (event, process and opinion). Statistically significant differences in consistency were found in regard to length but not in regard to the categories of topic. We argue that the expression inconsistency is a clear sign of a retrieval problem and that query expansion based on semantic relationships can significantly improve retrieval performance on free text sources.

## **1. Introduction**

Information Retrieval (IR) research has paid a lot of attention to text representation and query/document matching methods. Several text and discourse level phenomena are also well known in IR [1, 2]: synonymy, ambiguity (e.g., homonymy, polysemy), cohesion (references such as anaphora and ellipses). However, very little is known about how texts present the phenomena, objects, events or topics they discuss, how much variation there is. Consequently, we do not know whether such knowledge would be helpful for understanding IR phenomena, designing IR systems or using them for retrieval – performing queries.

Linguistic variation of document texts is among the greatest challenges to IR methods. The variation in presenting the subject matter greatly affects IR and IR system design. The variation of expression can be studied at the most specific level when there are two or more documents discussing precisely the same topic. News documents provide an easily available source for this purpose. News events are regularly reported through multiple competing news channels. There are several professional reporters writing to the same audiences about the same news phenomena, objects, events or topics (henceforth news events). Therefore many factors inducing variation are controlled. What remain are the personal styles of reporters in using the possibilities of natural language – of producing text. In this paper we study the variation of news reporting when the news articles report on the very same events in similar newspapers (in regard to language, size and type of the newspaper, as well as the size and type of the audience).

Another motivation for the study of news articles is the ubiquity of news, the constant demand for news, and their delivery through multiple competing news channels. In modern environments any event is likely to be reported in multiple versions in multiple news articles or video features.

There is a lot of redundancy in the broadcast and understanding the (in)consistency of reporting may help in dealing with news retrieval and filtering, either by utilizing or by controlling redundancy.

We are interested in the variation of expression in newspaper articles at the level of words and phrases. Most IR systems operate at the same level, treating words and phrases as strings for comparison. We are interested in the use of synonyms, word form variation, compounds vs. phrases, conceptual variation, etc. The main question is how consistent the articles are in their choice of words and phrases of expression when reporting on the same news events. We are also interested in the variation of expression at the conceptual level. High concept consistency may render semantic QE (for query expansion; e.g. [3, 4]) as a potential method. Thus our second question is how consistent the articles are in their choice of concepts when reporting on the same news events. Thirdly, we aim at reporting on the types of inconsistency.

Indexer and searcher consistency in selecting index/search keys from free text or a controlled vocabulary has been studied in the context of bibliographic information retrieval. Saracevic et al [5] have developed formulae for calculating the pairwise asymmetric consistency figures between actors (e.g., indexers or searchers). The consistency of a group of actors may then be calculated as the average of all pairwise asymmetric consistency figures of the group. Jones [6] and Cleverdon [7] have shown that inter-indexer consistency tends to be fairly low. Iivonen [8] showed that inter-searcher consistency at the search term level also was fairly low. However, after concept oriented normalization of expression (word form, synonym and hierarchy control), both inter- and intra-searcher consistency was at a high level (up to 93% for simple and specific requests in [8], using the formulae of [5]). In this paper we employ the consistency calculation methods to different newspaper articles by different reporters presenting the same news event.

This article is based on the research carried out by Lehtokangas [9]. It is organized as follows so that Section 2 presents the research problems, methods and data, and Section 3 the findings. Section 4 discusses the findings and Section 5 closes the paper.

## **2. Problems, methods and data**

The main question is how consistent newspaper articles are in their choice of words and phrases of expression when reporting on the same news events. The second question is how consistent the articles are in their choice of concepts when reporting on the same news events. Thirdly, does the consistency depend on the type of the articles/news events or on their length? Finally, what types of inconsistency are there between the articles?

Three categories of article type were identified: event, process and statement (of opinion). Events were characterized by a particular important moment after which the world was different. For example, an accident took place, a decision was made, or an agreement was achieved. The event itself had a short duration (or was reported as such by focusing on the outcome). Statements typically reported on a person's, organization's etc. opinion on one or more matters. The article reviewed, for example, a public talk. Articles that were not classified as events or statements were classified as process articles. They reported on long lasting (or unlimited) activities or totally lacked the time perspective (e.g., unemployment in Germany in January 1996).

Two categories of article length were utilized: news wires contained at most 100 words while reports contained at least 200 words<sup>1</sup>. According to Sormunen [10], the average article length in a large Finnish daily newspaper was about 230 words and the median about 150 words.

---

<sup>1</sup> All short items, here called news wires, did not originate from news agencies. 12% of them were produced by the newspapers themselves.

These two dimensions yield six article categories for analysis. Altogether 60 sets of three articles, articles in each set reporting on the same news event, were selected from three Finnish daily, large and general newspapers. The newspapers were *Aamulehti* (AL), *Helsingin Sanomat* (HS), and *Turun Sanomat* (TS) which are (by circulation) the three largest newspapers in Finland, HS being clearly the largest. There were 10 sets of three articles for each of the six categories. The articles / news events were purposely selected to match the category criteria. They cover the period of January to April in 1996. When a news event in one newspaper was reported in one article and was divided into two articles in another newspaper, both articles were included from the latter.

The consistency of reporting was analyzed both at the level of expressions and at the level of concepts the expressions represent. The former relate directly to IR techniques (matching of strings) while the latter to the deeper structures of articles, which may be consistent, even if the surface would be inconsistent.

Article triplets were the units of observation in the consistency analysis. Each triplet of articles was processed as follows. Each article was read, its essential concepts were recognized, the expressions for these concepts were marked and finally normalized into basic word forms (e.g., singular nominative for nouns, and first infinitive for verbs).<sup>2</sup> There is a fundamental problem in the identification of essential concepts pertaining to a news event: which concepts are essential, when to stop? Some concepts are clearly essential in telling the difference of the news event from any

---

<sup>2</sup> Note that Finnish is highly inflectional and without normalization expression consistency would be very low. Similar normalization is possible for database indexing and therefore the findings are relevant in the IR context (also for other languages).

other news event. Some others clearly refer to extraneous concepts, even to linguistic ones. There are no clear and general rules on how to tell the difference. In analyzing each article triplet, an understanding of the news event developed, suggesting the characterizing essential concepts. We can only argue for reasonable identification by the results of our reliability analysis. The reliability of concept recognition was assessed through partial repeated analysis of 36 articles (20% of the data; 6 times 2 sets of 3 articles) one year later. Concept recognition reliability was found to be high, about 88% for wires and 90% for reports. The expression sets, reflecting these concept sets and submitted to the consistency analysis, were to 95% (wires) and 98% (reports) the same. Both the original and the re-identification were done by the first author. Another analyst might have chosen the set of essential concepts differently for each article triplet. However, as long as the numbers of concepts per news event are at the same level, the findings of concept or expression consistency are not likely to be much affected. Had another analyst been more critical (or liberal) in accepting concepts as essential, the findings had shown higher (respectively, lower) consistency. However, this would not greatly affect the findings on the relationships of consistency at the concept and expression levels, nor the types of inconsistency discovered. Figure 1 illustrates the process of concept and expression identification.

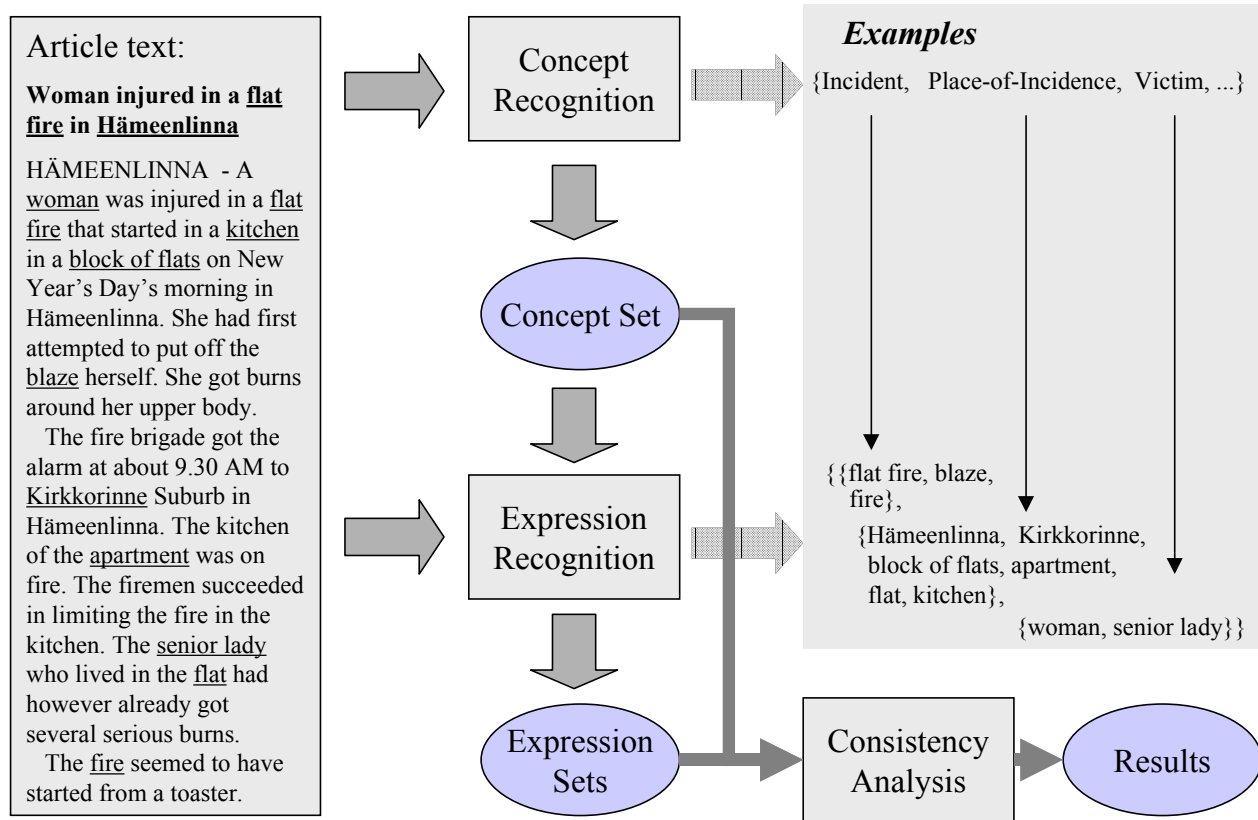


Figure 1. The process of concept and expression identification.

There was no list of concepts to aid in the identification of the essential concepts. Each triplet of articles created a concept set in a content driven way. Quite naturally, the concepts often contained typical news concept categories such as actor, action, object, place, time, reasons, etc. because that is the genre of news reporting [11, 12, 13].

Very little limitations were put on the possible expressions representing concepts. When identifying expressions for concepts, all expressions in the text contexts suggesting the concepts were collected. In subsequent classification these quite naturally fell into the categories of synonymous expressions, expressions at different hierarchical levels (generic or partitive), apposition attributes (e.g., lieutenant Kargbo) and associative relationships, as well as derivatives and spelling variants for compounds. Only some word categories were excluded because they are not used for



retrieval. Basically, expressions were identified as single words (which may be compounds in Finnish). However, proper names were identified as phrases. After word form normalization, concept and expression consistency were calculated.

The formula for the asymmetric consistency (applied from [5]) between articles 1 and 2 is:

$$CE_{1,2} = |E_1 \cap E_2| / |E_1|$$

where  $E_1$  and  $E_2$  denote the expression (or concept) sets of articles 1 and 2. The average (symmetric) consistency between the two articles is  $(CE_{1,2} + CE_{2,1}) / 2$ . For each set of three articles, the average consistency is the average of all pairwise consistencies. The consistency figure 100% means that all three articles agreed on all expressions (concepts), while the figure 0 % indicates that all articles had completely disjoint expressions (concepts). The consistency figures will be reported as averages in the six article categories and as the average of the whole data set. The statistical significance of expression (concept) consistency between articles was tested by the Chi-square test. The dependency between the consistency figures of the expression and concept levels was examined through the Pearson correlation coefficient.

### **3. Findings**

#### **3.1. The data**

The number of expressions and concepts recognized in the analysis is reported in Tables 1 and 2 by article type and length. The grand averages for expressions and concepts are 41.8 and 11.6 per article, respectively. The averages per article triplets were calculated by first taking the unions of the expression sets of articles in each triplet, then counting the number of expressions in each union, and finally averaging over the triplets, and similarly for concept sets. These figures reflect

the fact that our data contains over 4400 expressions for over 700 concepts for 60 news events (or article triplets). The differences between article and triplet averages suggest that there indeed are clear inconsistencies between the articles at the expression level and minor ones at the concept level. As expected, the data also reveal that there is a significant difference in the number of concepts (and their expressions) between the short news wires and longer reports. While the number of concepts per article does not vary by article type, statements have somewhat less expressions per article than the other two types.

Table 1. The average number of expressions for core concepts per article by article length and type (N=180 articles)

|                   |             | Article Type |         |           |         |             |
|-------------------|-------------|--------------|---------|-----------|---------|-------------|
|                   |             | Event        | Process | Statement | Average | Per triplet |
| Article<br>Length | Wire        | 17.9         | 20.3    | 17.8      | 18.7    | 23.8        |
|                   | Report      | 73.4         | 66.0    | 55.3      | 64.9    | 123.9       |
|                   | Average     | 45.6         | 43.2    | 36.6      | 41.8    | 73.8        |
|                   | Per triplet | 82.3         | 80.7    | 58.5      | 73.8    |             |

A comparison between tables 1 and 2 reveals that the average reduction factor from expressions to concepts was 6. In other words, on the average there were 6 expressions for each concept in our data.

Table 2. The average number of identified core concepts per article by article length and type (N=180 articles)

|                   |             | Article Type |         |           |         |             |
|-------------------|-------------|--------------|---------|-----------|---------|-------------|
|                   |             | Event        | Process | Statement | Average | Per triplet |
| Article<br>Length | Wire        | 8.0          | 8.1     | 9.0       | 8.4     | 8.7         |
|                   | Report      | 15.8         | 13.9    | 14.5      | 14.7    | 16.1        |
|                   | Average     | 11.9         | 11.0    | 11.8      | 11.6    | 12.4        |
|                   | Per triplet | 12.8         | 11.8    | 12.6      | 12.4    |             |

### 3.2. Expression consistency

Expression consistency by article length and type is reported in Table 3. The overall expression consistency is nearly 65%. Expression consistency does not greatly vary by article type but does vary, in each article type category, and in the whole data set, considerably by article length. Statements seem to be slightly more consistent than the other two types. Statements also contained slightly less expressions for concepts, which may have increased consistency. Also, statements may often be based on handouts and press releases, which are shared among the journalists. This would increase consistency.

Table 3. Expression consistency by article length and type (N=60 news events, 180 articles)

|         |         | Article Type |         |           |         |
|---------|---------|--------------|---------|-----------|---------|
|         |         | Event        | Process | Statement | Average |
| Article | Wire    | 77.8         | 83.7    | 87.0      | 82.8    |
| Length  | Report  | 46.7         | 40.4    | 53.2      | 46.8    |
|         | Average | 62.3         | 62.1    | 70.1      | 64.8    |

Short wires are fairly consistent — 78% to 87% with an average of 83 % — while longer reports are considerably less consistent in the range of 40% to 53%. This difference may be explained by the origin of the article texts: short wires are often produced by news agencies and printed in newspapers with minor or no modifications at all. Nevertheless, even for wires the consistency was clearly below 100%. This is a good baseline for comparing the consistency of reports. Longer reports are more likely to contain more editorial material and be based on the observations / research of the newspaper's own reporters. Longer reports also contain much more expressions thereby providing more chances for inconsistency than wires.

While the average consistency figures in Table 1 indicate a dependency on article length, there was great variation between the article triplets in each of the six categories. For example, consistency in report-length statements varied from 28 % to 86 %. The standard deviation of consistency was in the whole data set 23.4, for wires 14.8 and reports 15.0.

### 3.3. Concept consistency

Not surprisingly, the concept consistency figures reported in Table 4 are considerably higher than the expression consistency figures. The overall concept consistency is nearly 95 %. All variation is reduced to fairly narrow margins while article length still explains some of the variation in each article type category and in the whole data set. The standard deviation of consistency was in the whole data set only 5.4. Several concepts identified for the articles were fairly broad and could have been further divided into components. This would have increased the chances for inconsistency.

Table 4. Concept consistency by article length and type (N=60 news events, 180 articles)

|         |         | Article Type |         |           |         |
|---------|---------|--------------|---------|-----------|---------|
|         |         | Event        | Process | Statement | Average |
| Article | Wire    | 95.5         | 98.3    | 98.0      | 97.2    |
| Length  | Report  | 92.8         | 91.6    | 91.8      | 92.0    |
|         | Average | 94.1         | 94.9    | 94.9      | 94.6    |

These findings are important because the identified concepts are the essential concepts for each article / news event and thus the articles are likely to be relevant precisely for requests concerning those concepts.

### 3.4. Statistical significance

The dependency of expression (concept) consistency on article type and length was tested through the chi-square test. The test statistics revealed that, at both expression and concept levels,

consistency seems to depend on article length (significance level 0.01) but not on article type. This is quite obvious in Tables 3 and 4.

The correlation between expression consistency and concept consistency was calculated as the Pearson correlation, yielding the fairly high factor of 0.65 in the whole data set. Consequently, 42.8 % of the variation of expression consistency was explained by the variation of concept consistency. The rest was due to other factors, e.g., the reporters' choices between synonymous expressions.

### 3.5. Types of inconsistency

The types of inconsistency between the articles quite naturally fell into well-known categories of semantic relationships (synonymy, hierarchical and associative relationships) and into categories of lexical expression similarity (derivations, compounds). Below we discuss briefly these inconsistency types. The ISO standard [14] was used as a guideline for consolidating the classification of expression relationships.

*Synonymy* is one type of *semantic relationships* causing inconsistency. It is common in natural language texts [15] and is used, e.g., to avoid tautology. Between texts it may appear due to various cultural and idiosyncratic reasons.

The collision took place at 18.50 in Lahnajärvi. (HS)

The crash happened at 18.50 in Lahnajärvi about a kilometer toward Helsinki. (AL)

Below are examples on the types of synonymous expressions in the data [N.B. these English examples are similar to the Finnish ones in our data, but not always translations]:

*Abbreviation* - *Full expression:*

TV-company - television company

UN - United Nations

*Phrase* - *Compound:*

seat belt - seatbelt

jet lag - jetlag

*Foreign expression* - *National expression:*

gallup - opinion poll

*Transliteration*

Jihye Ayash - Yahya Ayyash - Jahja Aijash

*Hierarchical relationships.* Hyponyms were frequently used in the articles. For example, articles about an explosion in a flat referred to the flat by six different expressions at three levels of hierarchy (all expressions single words or compounds in Finnish, here indicated by hyphens):

Top: "apartment" (HS, AL, TS)

Mid: "flatlet" (HS, TS), "apartment-in-a-block-of-flats" (HS, TS),  
"explosion-flat" (HS)

Bottom: "first-floor-flat" (AL), "explosion-flatlet" (HS)

*Associative relationships.* Sometimes the articles used a slightly different approach to the news events and thus the expressions were related by association. For example, the articles reporting on president Clinton's speech on illegal immigrants used the expressions (illegal) "aliens" vs. (illegal) "immigration".

*Derivations* are the first type of *lexical expression similarity*. Derivatives are formed by adding prefixes or suffixes to root words. This is very productive and frequent in Finnish [16, 17]. For example:

- Both Yeltsin and Lukashenko stressed that the members of the union remain as independent states. (HS)
- Both parties will retain their independence, Yeltsin told to Russians and Belorussians. (AL, TS)

*Compounds* are very frequent in Finnish and they are formed quite freely [18]. Many are permanent, lexicalized while many others are just temporary. In many articles, the same concept was referred to by different compounds sharing the head word and the modifier expressing the viewpoint, for example all the following refer to “talo” (house) with different viewpoints:

- "kerrostalo" — a block of flats (AL, HS, TS)
- "vuokratalo" — a house with flats for rent (AL, HS, TS)
- "räjähdystalo" — a house where the explosion took place (HS)
- "elementtikerrostalo" — a block of flats built from elements (TS)

English readers should bear in mind that compounds are frequent in many European languages, e.g., the Nordic languages, German, Dutch, and Greek. The consistency phenomena are likely to be similar.

#### **4. Discussion**

In summary, the statistical findings are as follows:

- Expression level consistency in the data set was 65 %, on the average.



- Article length explained much of the expression level inconsistency (83 % for wires, 47 % for reports) although the variation within the categories was considerable.
- Concept level consistency in the data set was 95 %, on the average, with article length still explaining much of the remaining variation (97 % for wires, 92 % for reports) although there was variation within each of the six categories.

These findings are important because the identified concepts are the essential concepts for each article / news event and thus the articles are likely to be relevant precisely for requests concerning combinations of those concepts. They should be *highly relevant to each other*, i.e., if any of them was the query, the others should be top-ranking results.

The medium level expression consistency tells that the textual overlap between articles, based on the keys representing their essential concepts, is far from complete *even when the articles present the very same news event*. Further, expression consistency falls below 50 % when longer articles are considered. In spite of being highly relevant to each other, they were textually significantly different. This clearly reflects a retrieval problem. Whether or not they have a core set of concepts represented consistently, was not checked.

Articles reporting on the very same news events are excellent examples of *polyrepresentation* in human communication [19]. The article triplets present inconsistent expressions in categories of semantic relationships (synonymy, hierarchical and associative relationships) and of lexical expression similarity (derivations, compounds). It is not particularly striking that natural language presents phenomena we are familiar with in documentation languages. However, our findings suggest that, as information storage and retrieval increasingly moves away from services based on intellectual indexing, we still must take intellectual writing into account, in all its richness.

While this is not a novel finding, we have shown that there is a lot of variation of expression based on semantic and lexical inconsistency *even when news articles report on the very same news event*. Among topically relevant articles reporting on different news events, such variation is bound to be greater. Such variation may be accounted for by information retrieval methods utilizing semantic and lexical relationships.

Query expansion (QE, [20]) is a means to overcome surface level inconsistency of expression. The high conceptual consistency in our findings suggests that the problem of variation really lies in the expression. Kristensen [21], among others, found that the overlap between document sets retrieved through different expansions (synonyms, hierarchical relations, associations; Boolean retrieval) of the original search keys was fairly small. This suggests that expression level inconsistency / concept level consistency found in this study has clear retrieval effects that may be accounted for by a semantic query expansion approach. Later Kekäläinen [3, 22] found that concept based QE, together with concept structured queries, seems to yield the best performance also in probabilistic retrieval, and in retrieving highly relevant documents, in particular [23]. It remains, however, to be seen whether it is possible to develop and make generally available semantic sources for QE (or whether the cost is justifiable in terms of improvements). For example, the WordNet [24] has not so far provided clear-cut improvements in QE [25, 26].

Recent findings by Sormunen and others [4] indicate, from another angle, that article texts indeed present variation of expression based on semantic and lexical inconsistency. They compared the linguistic and conceptual features of highly relevant articles to those of the less relevant (but retrieved) ones for 18 requests (some 1200 articles at four levels of relevance). They found out that highly relevant articles, as opposed to less relevant articles:

- (a) discussed the request topic at more length
- (b) dealt with more aspects of the request
- (c) contained more expressions pertaining to the request
- (d) provided more unique expressions to refer to the concepts of the request.

The more relevant the articles were, the higher was the probability that they contained at least one expression for each facet (b), suggesting higher conceptual consistency with the request. However, (d) allows for higher expression inconsistency between the articles and the request (and with each other). Sormunen and others [4] also found that concept-based query expansion, with suitable query structures, improved the retrieval of highly relevant documents.

Our findings are based on the article texts and on analyzing, for each news event, the most essential concepts and their expressions. Whether some concept / expression could be expected to be used in an IR request (or relevant to such a request) was not taken into account. For example, a slalom skier was also called by attributes like "wildhead", "ace" and "fighter" in addition to his proper name. Whether such words are used consistently or not, is not serious for IR.

Large events, which last a long time, naturally produce multiple articles with multiple approaches, and thus there is lots of variation between them, of course. Similarly, topical requests, not focusing on a single event, result in article sets that also present more variation than single-event articles. These facts in no way undermine our findings, quite the contrary.

We do not know whether the phenomena studied here may be identified in text corpora of other discourses, e.g., whether (or to what degree) document texts in various academic or professional domains are more consistent in expression than newspaper articles. Nevertheless, the news domain is a general and important one, and tends to produce several articles on each news event.

Many of them are simply extraneous and may still occupy high ranked positions in retrieval (or pass the filter *in corpore* in filtering). If this is to be controlled, plain expression level methods may not prove sufficient due to variation of expression. However, if the concepts represented, rather than expressions, are identified, the chances are much better.

Our data does not allow us to conclude, whether or not a wire-length article would be likely to provide a subset of the essential concepts of a report-length article on the same news event. While this is not an unreasonable hypothesis, it should be tested.

## **5. Conclusions**

We investigated the consistency of newspapers in their choice of words and concepts when writing about the same news event. News articles on the same news events were taken from three Finnish newspapers and compared in regard to their central concepts and expressions representing the concepts in the news texts. Consistency figures were calculated for each set of three articles (the total number of sets was 60). Inconsistency in expressions and concepts was found between news articles from different newspapers. The mean value of consistency calculated on the basis of expressions was 65 %; this however depended on the article length. For short news wires consistency was 83% while for long articles it was only 47%. At the concept level, consistency was considerably higher, 95 %, ranging from 92 % to 97% between short and long articles. The news articles also represented three categories of topic (event, process and opinion). Statistically significant differences in consistency were found in regard to length but not in regard to the categories of topic.

These findings suggest that IR might benefit from QE based on semantic relationships and that such expansion can be employed effectively by suitable query formulation. Moreover, recogniz

ing conceptual consistency (similarity) between articles might help identifying extraneous articles in news filtering and retrieval, and help lumping them together for presentation. More generally, our findings call for attention on document texts. Document and request characteristics might provide new possibilities for systems development and certainly new insight on, and understanding of, the processes of IR. A sufficient theory of information retrieval should take, in addition to statistical properties of document texts, also their linguistic and conceptual properties into account as documents are purposive units of human communication.

### **Acknowledgements**

We thank Dr. Jaana Kekäläinen for her constructive comments on earlier versions of the manuscript. This research is part of the research project *Query structures and dictionaries as tools in concept-based and cross-lingual information retrieval* funded by the Academy of Finland (Research Projects 44704; 49157). This research was completed while the second author was on a leave at the Department of Information Studies, University of Sheffield, UK, fall 2000.

### **References**

1. Haas, S.W. Natural language processing: toward large-scale, robust systems. In: *Annual Review of Information Science and Technology*, 31. Medford, New Jersey: [Published by] Information Today, Inc., [for the] American Society for Information Science, 1996, 83-119.
2. Pirkola, A. *Studies on linguistic problems and methods in text retrieval*. Tampere, Finland: University of Tampere, Acta Universitatis Tamperensis 672, 1999. Ph.D. dissertation.
3. Kekäläinen, J. *The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval*. Tampere, Finland: University of Tampere, Acta Univer

sitatis Tamperensis 678, 1999. Also available at:

[http://www.info.uta.fi/research/postscript\\_docs/JK1\\_99.pdf](http://www.info.uta.fi/research/postscript_docs/JK1_99.pdf) (visited 16 November, 2000).

Ph.D. dissertation.

4. Sormunen, E., Kekäläinen, J., Koivisto, J. and Järvelin, K. Document text characteristics affect the ranking of the most relevant documents by expanded structured queries. *Journal of Documentation*, 57, 2001, accepted for publication.
5. Saracevic, T. and al. A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science*, 39(3), 1988, 161-176.
6. Jones, K.P. How do we index? A report of some Aslib informatics group activity. *Journal of Documentation*, 39(1), 1983, 1-23.
7. Cleverdon, C. Optimizing convenient online access to bibliographic databases. *Information Services & Use*, 4(1/2), 1984, 37-47.
8. Iivonen, M. Consistency in the selection of search concepts and search terms. *Information Processing and Management*, 31(2), 1995, 173-190.
9. Lehtokangas, R. *Sananvalinnan yhdenmukaisuus sanomalehtiutisissa* [Expression consistency in newspaper articles]. Tampere, Finland: University of Tampere, Department of Information Studies, M.Sc. Thesis, 1999. (In Finnish)
10. Sormunen, E. *A Method for measuring Wide Range Performance of Boolean Queries in Full-Text Databases*. Tampere, Finland: University of Tampere, Acta Electronica Universitatis Tamperensis, 2000. Also available at: <http://acta.uta.fi/pdf/951-44-4732-8.pdf> (visited 16 November, 2000). Ph.D. dissertation.

11. van Dijk, T.A. Structures of news in the press. In: van Dijk, T.A., ed. *Discourse and communication: New approaches to the analysis of mass media discourse and communication*. Berlin: Gruyter, 1985, 69-93.
12. van Dijk, T.A. *News analysis: case studies of international and national news in the press*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
13. van Dijk, T.A. *News as discourse*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
14. *ISO International Standard 2788. Documentation - Guidelines for the establishment and development of monolingual thesauri*. Paris: International Organization for Standardization, 1986.
15. Sivula, J. Synonyymit [Synonyms]. Vesikansa, J., ed. *Nykysuomen sanavarat*. Helsinki, Finland: WSOY, 1989, 183-199.
16. Lepäsmä, A.-L., Lieko, A. and Silfverberg, L. *Miten sanoja johdetaan: suomen kielen johtopöytä* [How are words derived: A textbook in derivation in Finnish]. Helsinki: Finn Lectura, 1996.
17. Vesikansa, J. *Johdokset* [Derivations]. Helsinki, Finland: WSOY, 1978.
18. Häkkinen, K. *Mistä sanat tulevat: suomalaista etymologiaa* [Where do the words come from: Finnish etymology]. Helsinki: Suomalaisen kirjallisuuden seura, 1990.
19. Ingwersen, P. Cognitive perspectives of information retrieval interaction. *Journal of Documentation*, 52(1), 1996, 3-50.

20. Efthimiadis, E.N. Query expansion. In: *Annual Review of Information Science and Technology*, 31. Medford, New Jersey: [Published by] Information Today, Inc [for the] American Society for Information Science, 1996, 121-187.
21. Kristensen, J. Expanding end-users' query statements for free text searching with a search-aid thesaurus. *Information Processing & Management*, 29(6), 1993, 733-744.
22. Kekäläinen, J. and Järvelin, K. The co-effects of query structure and expansion on retrieval performance in probabilistic text retrieval. *Information Retrieval*, 1(4), 2000, 329-344.
23. Järvelin, K. and Kekäläinen, J. IR evaluation methods for retrieving highly relevant documents. In: Belkin, N., Ingwersen, P. and Leong, M-K., eds. *Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR '00)*. New York, NY: ACM Press, 2000, 41-48.
24. Miller, G. A. WordNet: A lexical database for English. *Communications of the ACM*, 38(11), 1995, 39-41.
25. Crestani, F., Sanderson, M., Theophylactou, M. and Lalmas, M. Short queries, natural language and spoken document retrieval: Experiments at Glasgow University. In: Voorhees, E.M. and Harman, D.K., eds. *Information technology: The Sixth Text Retrieval Conference (TREC-6)*. Gaithersburg, MD: National Institute of Standards and Technology, 1997, 667-686.
26. Voorhees, E. Query expansion using lexical-semantic relations. In: Croft, W.B. & van Rijsbergen, C.J., eds. *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. New York, NY: ACM, 1994, 61-69.