Ord i Dag: Mining Norwegian Daily Newswire

Unni Cathrine Eiken[*], Anja Therese Liseth[*], Hans Friedrich Witschel[+], Matthias Richter[+] and Chris Biemann[+]

[*]University of Bergen, AKSIS
Allégaten 27, 5007 Bergen, Norway
unni@eiken.no, anjaliseth@gmail.com

[+]University of Leipzig, NLP Department
Augustusplatz 10/11, 04109 Leipzig, Germany
{witschel, mrichter, biem}@informatik.uni-leipzig.de

Abstract. We present *Ord i Dag*, a new service that displays today's most important keywords. These are extracted fully automatically from Norwegian online newspapers. Describing the complete process, we provide an entirely disclosed method for media monitoring and news summarization.
For keyword extraction, a reference corpus serves as background about average language use, which is contrasted with the current day's word frequencies. Having detected the most prominent keywords of a day, we introduce several ways of grouping and displaying them in intuitive ways. A discussion about possible applications concludes.
Up to now, the service is available for Norwegian and German. As only some shallow language-specific processing is needed, it can easily be set up for other languages.

1 Introduction

Machine aided media monitoring is a service that has been offered commercially for years, typically covering thousands of channels and providing products from keyword monitoring to media resonance analysis. *Ord i Dag* (Word of the day) is a new way of monitoring news, which is interesting for academic research as well as for the layman. By using different ways of presenting today's most important keywords, we prepare a good overview of what the media considers as today's most interesting news and topics. We also present a method of monitoring these events over time.

1.1 Motivation

Media produce large and ever growing amounts of content on a regular basis. Texts account for a significant part thereof and its analysis constitutes a promising field of research. Due to the amount of data it is an obvious idea to bring into the field statistical methods which can help to single out new and interesting events and topics.

The criteria used in commercial solutions are, if at all, often not fully laid open. *Ord i dag*, as described in this work, is a fully disclosed selection and presentation process, providing a solid data foundation for research on relations and developments.

1.2 Related Work

In recent years, topics such as summarization, clustering, filtering and tracking of news have been covered, often related to the novelty track in TREC and TDT [1]. In 1997, the Altavista search engine featured *LiveTopics* (see `http://www.samizdat.com/script/lt1.htm`), which included a graph calculated from words in current news. [2] describes *Newsblaster*, which groups stories by a Topic Detection and Tracking system and generates multi document summaries for news on a daily basis. This is comparable to approaches by *Google News* and others, but with *Newsblaster* keeping an index of past days. Another multi document summary centered approach is *NewsInEssence* [3]. Visualization interfaces for news search sites have been built and made available on the Web: *Newsmap* at `http://www.marumushi.com/apps/newsmap/` presents the groups and stories from *Google News* in form of a treemap [4], which represents the amount of coverage of topics in differently sized labeled boxes. *In the News* at `http://news.stamen.com/` combines different data sources to display a real-time overview on news based on bar diagrams and featuring sparklines [5] for monitoring change over time. Sparklines are also used for the visual display of frequency information at the *Wörter der Woche* calculated from each issue of the German newspaper "Die Zeit" by the Berlin-Brandenburg Academy of Science. These works suggest that there is a niche for news analysis and visualization in the recently growing field of Visual Analytics [6].

1.3 Outline

In this work we cover the full process of obtaining a daily amount of 100-150 keywords that describe the most important events in the daily news. The process is split into two parts: the preparation of a reference corpus, which is outlined in section 2, and the daily processing of news data as explained in detail in section 3.

Section 4 deals with the presentation of the data: grouping related keywords for a user-friendly website layout. Visualization of daily keywords as well as time-dependent changes is exemplified. Section 5 concludes with discussing some possible extensions and applications.

All data is available in human- and machine-readable format on `http://wortschatz.uni-leipzig.de/wdtno/`.

2 Preparing the Reference Corpus

Figure 1 depicts the steps undertaken in order to prepare the reference corpus. A reference corpus is a large corpus that is used as a model for 'average' language use in the following.

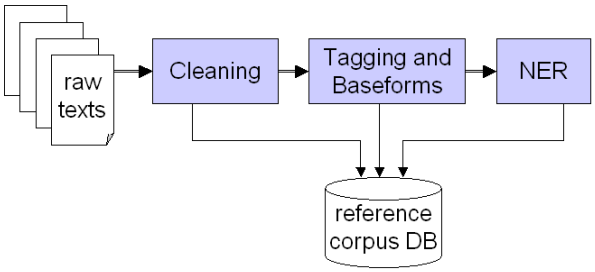

Figure 1: Process chain for preparing the reference corpus database

2.1 Preprocessing

For the purposes of this project we make use of the Norwegian newspaper corpus Norsk aviskorpus (`http://www.avis.uib.no`, cf. [7]), which is collected by the University of Bergen and consists of approximately 440 million running words from 1.3 million texts. It is augmented every day by roughly 200,000 words from a selection of Norwegian newspapers on the internet.

When transforming the corpus into our reference corpus, some cleaning was needed, i.e. stripping of unprocessed HTML tags, broken picture headings and external links. Although the everyday routine for collecting the newspaper texts is updated constantly, a large amount of commercials, names of journalists and photographers etc. had to be removed from older material. It was also necessary to set up routines to remove such unwanted parts, should they ever re-occur in new texts.

2.2 Linguistic Processing

Tagging and base form reduction. Part-of-speech tagging and base form reduction are important steps in the linguistic pre-processing. Tagging provides information on word classes; base form reduction maps several inflected forms onto one base form. Both actions are important for linguistic filtering as elaborated in section 3.1.

For Norwegian, there exists the Oslo-Bergen-tagger, a high-quality constraint-based tagger [8] that does not only assign word class information, but also is able to annotate base forms and syntactic functions. However, this rule-based tagger is too slow for granting topicality and its output provides much more information than needed for our purpose. We therefore re-engineered the tagger in a simple way: We

tagged a portion of the corpus and used the obtained triples (word, tag, base form) as training data for Compact Patricia Tree (CPT) classifiers.

Using the implementation from [9], CPT classifiers are trained to return a class, given a string. The number of classes is not restricted and the training set is perfectly reproduced. Due to the compact representation and an efficient search mechanism in the tree, CPTs can be used as lexical components for millions of words. The most important feature of CPTs, however, is their ability to generalise, i.e. to return classification guesses for unseen strings. For example, if an yet unseen word like *deministration* is classified, its class will be guessed based on training words with a longest common affix, e.g. *administration*. The same class will be assigned for similar strings. If several training words of different classes match with the same longest common affix, a class distribution is returned. CPTs can be trained on beginnings or endings of strings.

How to employ these classifiers for tagging and base form reduction is described in the following subsections.

Tagging with CPTs. As we need only rudimental word class information for filtering, we reduced the tag set of the Oslo-Bergen-tagger to the following basic categories: noun (N), verb (V), adjective (A), adverb (AV), cardinality (C), interpunctuation mark (IM) and others (S). For the most frequent 100,000 word forms of our tagged part of the corpus, we trained two CPT classifiers on these classes: one for prefixes and one for suffixes. Here, we only allowed one possible tag per string: POS-ambiguous words receive their most frequent tag, which is sufficient for our filtering task. With these top frequency words, we achieve a text coverage of about 96.3%. For these words, our classifiers yield unique classes that form the tags. For unknown words, the intersection of the two class distributions determines the tag.

This implements a unigram part-of-speech tagger that clearly does not meet the requirements of a full-fledged POS-tagger but is sufficient for the subsequent steps. A higher quality would be obtained when using e.g. a HMM tagger.

Base Form Reduction with CPTs. Unlike the low level POS-tagger, the quality of base form reduction with CPTs meets state-of-the-art requirements. Given a list of pairs (word, base form), reduction rules for the conversion from full form to base form are computed. Table 1 gives examples for the verb "stå" (to stand).

full form	stå	stående	står	stås	stått	sto	stod
reduction rule	0	4	1	1	2	1å	2å

Table 1: Reduction rules for some full forms of "stå"

The reduction rules consist of two parts: A number indicating how many characters should be cut from the suffix of the full form, and an optional string that is attached after the cut operation. We learn reduction rules as they are similar for words with the same inflection behavior.

For base form reduction, three CPTs are trained on suffixes: one for each open word class (nouns, verbs, adjectives). The tag from the unigram tagger is used for the classification.

Identification of Named Entities. The identification of named entities is a further important step in the linguistic processing of the data collection. In order to present users with an informative list of daily keywords, we need a way to recognize the multi word units that constitute named entities. We implemented a weakly supervised method described in [10] to recognize person names in the corpus. The algorithm takes very little input knowledge and performs iterative learning on unlabeled data. By drawing advantage of the fact that named entities display a high degree of regularity in their form, the algorithm bootstraps new instances of named entities based on a small set of initial names and classification rules.

The algorithm was initially supplied with a list of a few hundred common name elements, labelled as first or last names, a short list of common titles, and a set of classification rules and extraction patterns. The classification rules specify patterns that determine which tag a new name should be assigned. For example the rule TIT CAP* LN -> FN would entail that a capitalized word found between a known title and a known last name would be tagged as a first name. To ensure high classification precision, this decision is verified on other occurrences of this word.

By these means we construct a large, corpus-specific list of name parts that is used in a heavily gazetter-based Named Entity recognizer. In this NER component, the recognition of yet unseen names is carried out by the same classification rules as mentioned above (including the verification step). Using this approach, we collected a list of about 300,000 named entities from the corpus. These are used as multi word units in the following. The list includes names and titles, allowing entries such as *Jens Stoltenberg* and *statsminister Jens Stoltenberg*.

3 Processing the daily data

When processing daily data, the same preprocessing steps as for the reference corpus are carried out. Additionally, the reference corpus is used as a means of comparison for obtaining the words that are most prominent on that specific date. Figure 2 depicts the process chain.

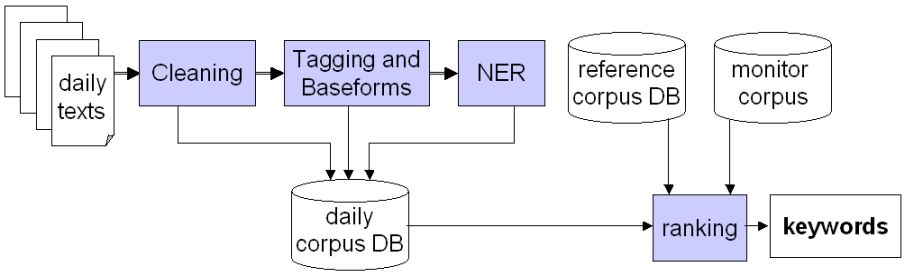

Figure 2: Process chain for obtaining keywords from daily corpus

3.1 Extraction of Keywords

Selection Criteria. Keywords are selected following the set of assumptions proposed in [11] with an additional language specific component and a monitor corpus. Concepts can be represented best by nouns and named entities. Therefore only these word classes are considered eligible for keyword selection. There are three groups of keywords that are treated differently: words present in the reference corpus, multi words and words not present in the reference corpus.

For words seen before, a difference analysis [12] based on the Poisson measure of surprise combined with a comparison of relative frequencies is carried out. A word is considered a keyword if its Poisson significance exceeds a threshold and if it occurs with sufficient frequency. The words not covered by the reference corpus become keywords if they can pass the normal frequency threshold test.

The static reference corpus may not cover recent developments. This shortcoming is addressed by using frequency data from the recent past as a monitor corpus. An *inclusion* rule selects keywords if they occur much more often on the present day than in the preceding five week frame. An *exclusion* rule is applied to drop words if on the current day they occurred relatively less often than in the week before.

3.2 Grouping of Keywords

In order to provide a usable overview of keywords for the news topics of the day, these need to be grouped. Two alternative grouping approaches are described below.

Keyword Categorization. The reference corpus contains source URLs for each article. These can be exploited for learning a classification with only a very small amount of human interaction. The preparation process consists of the following steps:

1. Extract all possible fragments of information from the source URLs that contribute to more than 10.000 sentences. This gave 145 category candidates.
2. By eliminating in a further step senseless strings and grouping differently written variations, the number of categories for display could be reduced to only 11: “Regional” (regional), “Innenriks” (home affairs), “Utenriks” (foreign affairs), “Politikk” (politics), “Økonomi” (economy), “Kultur” (culture), “Sport” (sports), “Utdanning” (education and research), “Bil” (automobile), “Forbruker” (consumer) and “Teknologi” (technology).
3. For each pair of word and category, a dependency statistics was calculated in the form of a likelihood ratio which gives high significance to words that – in the given category – occur with higher frequency and negative values to words which appear more rarely than predicted by the entire training collection.

In the daily process, new text can be classified by assigning for each word in each sentence a category weight, which takes into account both positive and negative values for all categories and the frequency of the word in the reference corpus. An assignment of the sentence to a category is made if the sum passes a threshold, which

has been set up to provide an average of 90% precision and 50 % recall in an evaluation on the training set with 10-fold cross-validation.
For texts, the sum of classifications of all sentences is calculated and the highest rated category is assigned to the text.
Clear advantages of this simplistic approach are that (except for setting up the categories) the process works fully automatically and is based on a wide coverage of vocabulary, which means that there are many features and thus enough support for reliable classification.

Candidate Clustering. Another form of organizing the *ord i dag*-keywords aims at a more granular and flexible presentation: instead of using a fixed set of categories, it might be interesting to model events using a constantly changing and more fine-grained classification that gives a quick overview of the day's events. This can be achieved by clustering the keywords and learning headlines for clusters. The headlines will be the new set of categories. As an example of why this is desirable, consider the categories and clusters presented in figure 3: The headline "Israel" in figure 3b) tells that something interesting has happened in Israel, whereas this fact is lost in figure 3a) within the general category "Utenriks" (foreign affairs).

However, a serious drawback of the cluster representation is the manual effort which it requires: headlines are assigned manually in the beginning and subsequent automatic headline assignments must be supervised.
Following is a detailed description of how we arrive at clusters and headlines:

1. *Feature Selection:* in order to describe a keyword, it is assigned a feature vector derived from its example sentences S, i.e. all of today's sentences in which it occurs. The feature vector consists of all other keywords that appear in S, weighted with their frequency.
2. *Clustering:* Now the keywords are clustered using their feature vectors via K-means with cosine as similarity measure
3. *Headline assignment:* when the process is carried out for the first time, all headlines have to be assigned manually to clusters. For each cluster which is labeled, its headline will be stored, together with a centroid of the cluster members' feature vectors.
4. *Inheriting headlines:* For the following days, the set of centroids of past clusters, together with their headlines are treated as categories and a new cluster C is classified using a Rocchio method, i.e. a centroid is computed for C and it inherits the headline from the cluster whose centroid is closest (most similar) to C's. This is only done if the similarity exceeds a threshold (currently 0.3).

Experience has shown that the system learns rather quickly, i.e. that after a few days a substantial part of the clusters receive an automatically assigned headline. The possibility to assign new headlines to newly emerging topics (i.e. to create new categories) is one of the strengths of this approach. However – since the classifier is not perfectly accurate – some "automatic" headlines need to be corrected manually. The advantage of this procedure – when compared to completely automatic headline assignment of any sort – is the fact that we can have a high level of abstraction (as can only be achieved by humans) while maintaining maximum flexibility (e.g. to invent categories when new topics emerge).

4 Presentation

4.1 Textual web interface

The textual web interface is split into two views: a *daily overview* and a *detailed word view* which are linked to each other via the term, respectively the date.

Overview. In the overview (figure 3a), the list of selected words is presented in alphabetical order for each of the categories, or alternatively, the cluster headlines (3b). Clicking on a word leads to its detailed view. The font size in the category view denotes the weight: the most important keywords can be spotted at a glance.

(a)

Clusters of the day: 15-03-2006

Google	Eustace · Fast · Google · Yahoo ·
Israel	Ahmed · fengsel · israeler · Jeriko · Mahmoud Abbas · observatør · Saadat ·
Kultur	Alnæs · forfatter · forlag · Gyldendal · Karsten Alnæs · Karsten ·
Kultur	Ullmann · Linn · kunst ·
Nyheter - innenriks	bensinstasjon · Bjerkeland · Bård Bjerkeland · Bård · hagle · Kjeller · Lillestrøm · politimann · Rösler · skytedrama · Statoil-stasjon ·
Nyheter - innenriks	bakterie · Folkehelseinstitutt · Forbrukerråd · Fosen · Froland · Gilde · Joakim Lystad · kjøttdeig · kjøtt · Lidl · Lystad · Mattilsynet · råd · Rudshøgda · slakteri · smitte · smittekilde ·
Nyheter - innenriks	Dalby · Jenta · nyresvikt · Oppland ·
Politikk - innenriks	barnehageplass · Djupedal · kunnskapsminister Øystein Djupedal ·
Sport	Bjørnar · Iditarod · Jeff · King · Mountain · Nome · sjekkpunkt · White ·
Sport - ski	Beckie · Bjørgen · Ella · Guro · Guro Strøm Solli · Scott ·
Sport - ski	Aamodt · Aksel · Aksel Lund Svindal · Antoine Deneriaz · Deneriaz · Kjetil André Aamodt · Kjus · Paralympics · Peter Fill · Åre · Schjatvet · Svindal · verdenscup ·
Sport - ski	Björn · Lind · Rønning ·
Teknologi - spill	Kutaragi · PlayStation · Sony ·

««14.03.2006««Words of the Day[overview]»»16.03.2006»»

(b)

Figure 3: Ord i dag of 15th of March, 2006 in a) category and b) cluster view

This weight is the ratio of relative frequency in the daily corpus compared to the reference corpus, called "weirdness index" in [13]. The resulting values are scaled logarithmically to font sizes of 50-200%. Furthermore, the displayed terms do not only differ in size but also in their lightness value: the light end side of the scale is used for small degrees of certainty in the classification and the dark end side for almost sure classifications.

Note that not all terms from the category view appear in the cluster view. This may be because they are contained in a cluster which has not been assigned a headline or in a cluster which is too small (≤ 2 elements) to be displayed.

Detailed word view. The detailed word view depicts information centered around one focus word. It consists of four parts as shown in figures 4 and 5.

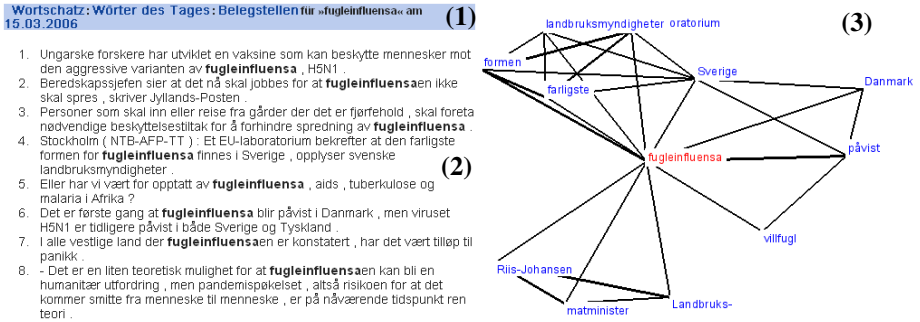

Figure 4: Ord i Dag detailed word view: usage and co-occurrence graph for *fugleinfluensa* (bird flu) on the 15th of March 2006.

First there are links back to the daily overview and pages explaining the process of selection and the collected materials (1). Then for the focus word and each of its forms a full list of occurrences in the daily corpus is given with the focus word emphasized in bold font style (2). For each sentence, a source reference contains a backlink to the original newswire article. The last two elements in the detailed view are an association graph (3) of co-occurrences and a combined frequency and co-occurrence graph for the focus word and the top co-occurrences of the focus word (as shown in figure 5). In the following section, these graphs are explained in detail.

4.2 Graphical Interface

Association Graph. For the association graph a selected fixed size set of sentence based co-occurrences is retrieved according to the co-occurrences' likelihood ratio [14]. As an additional constraint it is required for each node to have edges with at least the focus word and one more word from the set. The resulting graph is laid out fully automatically using simulated annealing as described in [15], helping the user to rapidly gain an overview of correlated terms.

Frequency / Co-occurrence Graph

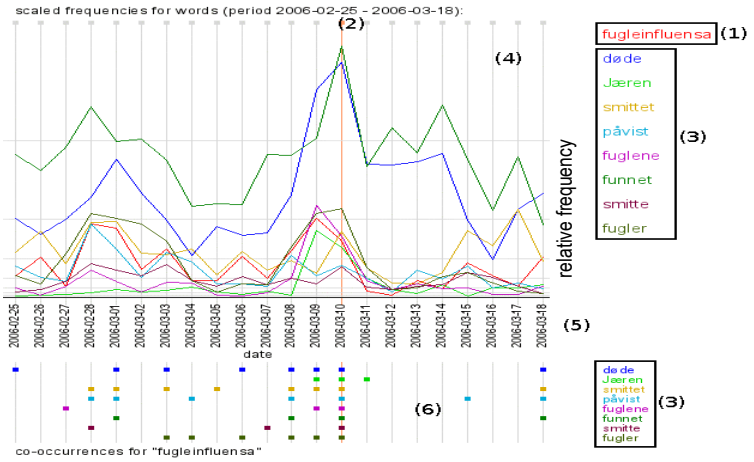

Figure 5: Frequency / co-occurrence graph displaying the focus word *fugleinfluensa* (1) for a specific date (2) with its eight most significant co-occurrences (3). The frequency graph (4) displays the time-dependent development of those words' relative frequencies in logarithmic scale on a date line (5). The co-occurrence matrix (6) displays whether joint peaks are significant.

4.3 Machine Readable Output

The selected terms are also made accessible in machine readable format as one RSS 2.0 (Really Simple Syndication) feed per category for use with RSS readers and for content syndication. The feeds contain a list of the words of the current day, their frequency and links to the respective pages of the web interface. The RSS feeds are linked from the overview page of the *Ord i dag* in a way that they are automatically offered to an RSS autodiscovery enabled web browser such as Mozilla Firefox. As a fallback the feeds are also listed on an overview page at `http://wortschatz.uni-leipzig.de/wdtno/RSS/`.

The RSS-feeds of the German version of *Wörter des Tages*, available from `http://wortschatz.uni-leipzig.de/wort-des-tages/`, are used by the publishing house Langenscheidt to get a reasonably sized and up-to-date selection of words that are proposed to learners of English on their website.

5 Conclusion and further work

The implementation of *Ord i dag* shows the language independency of the framework developed by the Wortschatz project (`http://www.wortschatz.uni-leipzig.de`, [16]). Although a range of language- and data source specific amendments had to be carried out for the Norwegian version, the original framework

is for the most part implemented directly with the Norwegian newspaper corpus as data source. Similar implementations could be carried out quite easily for other languages, provided the existence of sufficient corpus resources that are updated on a daily basis. But much more important than the potential of implementations for further and similar languages, the applications that now exist for Norwegian data can be used for a diversity of language related research. Some of these branches of research will be outlined in the remainder.

5.1 Neologisms

As the corpus consists of daily collected texts from newspapers on the web, it offers a convenient opportunity to monitor the rise and decline of words. The corpus can be consulted for information on when a word is used for the first time, how the frequency of use increases or decreases over time and ultimately when a word ceases being in common use. This is interesting on the one hand from a diachronic linguist's point of view; how long does it take before a new noun or name, such as *Google,* is in use as a verb, such as *to google*, or an adjective, such as *googled information*? On the other hand, this information can be of practical use as well, by providing a means of easy and fast creation of updated dictionaries of neologisms, or new words. But not only new singular terms, also new combinations of existing terms, be it multi-word units or simply new associations, can be tracked.

5.2 Trend monitoring

A further application is trend monitoring, a task that has received increased commercial interest in recent years (see: [17], part III). By consulting the corpus we can for instance see how long a certain case is covered by the media. How long does it take before the newspapers stop writing about a particular case? Do certain cases re-appear in the media after a time? Are there foreseeable time intervals between the re-occurrences of such cases? Through analyses of co-occurrences, we can also say something about the co-dependency of the media profiling of cases; do certain cases trigger the emergence of other cases? These analyses diverge from traditional commercial media monitoring in one important aspect: rather than monitoring a set of predefined keywords, we can monitor on a more objective basis, essentially monitoring *language use* as well as *media*. The information obtainable through a media monitoring analysis of corpus data is of interest for several fields of research, ranging from linguistics, via humanities, to economic fields.

References

1. Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation, Wayne, C., In Proceedings of LREC (2000) 1487-1494.

2. McKeown, K. R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans J. L., Sable, C., Schiffman, B., and Sigelman, S.: Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. In Proc. of the Human Language Technology Conference (2002)
3. Radev, D., Otterbacher, J., Winkel A., Blair-Goldenson, A.: NewsInEssence: Summarizing Online News Topics. Communications of the ACM. Vol. 48, No. 10, (2005) 95-98
4. Bederson, B.B., Shneiderman, B., and Wattenberg, M.: Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies. ACM Transactions on Graphics (TOG), 21, (4), (2002) 833-854.
5. Tufte, E.: Beautiful Evidence. To appear. Draft at: http://www.edwardtufte.com /bboard/q-and-a-fetch-msg?msg_id=0001OR&topic_id=1 (2006)
6. Thomas, J. J. and Cook, K.A. (eds.): Illuminating the Path: The Research and Development Agenda for Visual Analytics. IEEE Press (2005)
7. Hofland, K.: A Self-Expanding Corpus Based on Newspapers on the Web. Proceedings of the Second International Language Resources and Evaluation Conference. Paris: ELRA (2000)
8. Johannessen, J.-B., Hagen, K. and Nøklestad, A.: A Constraint-based tagger for Norwegian. In 17th Scandinavian Conference of Linguistics, Odense Working Papers in Language and Communication 19, University of Southern Denmark, Odense, Vol. 1, (2000) 31-47
9. Witschel, H.F. and Biemann, C.: Rigorous dimensionality reduction through linguistically motivated feature selection for text categorisation. Proceedings of NODALIDA, Joensuu, Finland (2005)
10. Quasthoff, U., Biemann, C., Wolff, C.: Named Entity Learning and Verification: Expectation Maximisation in Large Corpora, Proceedings of CoNNL-2002, Taipei, Taiwan (2002) 8-14
11. Richter, M.: Analysis and Visualization for Daily Newspaper Corpora. Proceedings of RANLP, (2005) 424-428
12. Faulstich, L., Quasthoff, U., Schmidt, F., Wolff, C.: Concept Extractor - Ein flexibler und domänen-spezifischer Web Service zur Beschlagwortung von Texten. In Proceedings of ISI 2002, Regensburg (2002)
13. Ahmad, K., Gillam, L., Tostevin, L.: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). In Proceedings of TREC-8. Washington: National Institute of Standards and Technology. (2000) 717-724
14. Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19:1. (1993)
15. Davidson, R. and Harel, D.: Drawing graphs nicely using simulated annealing. ACM Transactions on Graphics, 15(4), (1996) 301–331
16. Biemann, Chr., Bordag, S., Heyer, G., Quasthoff, U., Wolff, Chr.: Language-independent Methods for Compiling Monolingual Lexical Data, Proceedings of CicLING 2004, Seoul, Korea and Springer LNCS 2945, Springer (2004) 215-228
17. Berry, M.W.: Survey of Text Mining: Clustering, Classification and Retrieval. Springer (2003)