Mining Textual Contents of Quarterly Reports

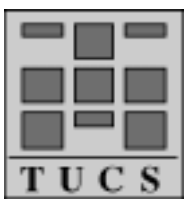
Antonina Kloptchenko
Turku Centre for Computer Science, Institute for Advanced Management System Research, Åbo Akademi University, Lemminkäisenkatu 14, FIN-20520 Turku, Finland

Camilla Magnusson
Department of General Linguistics, University of Helsinki, Helsinki, Finland

Barbro Back
Turku Centre for Computer Science, Institute for Advanced Management System Research, Åbo Akademi University, Lemminkäisenkatu 14, FIN-20520 Turku, Finland

Ari Visa
Tampere University of Technology, Department of Information Technology, Tampere, Finland

Hannu Vanharanta
Pori School of Technology and Economics, Pori, Finland



Turku Centre for Computer Science
TUCS Technical Report No 515
May 2002
ISBN 952-12-1138-5
ISSN 1239-1891

Abstract

A huge amount of electronic information concerning companies' financial performance is available in organizational databases and on the Internet today. Numeric financial information is important for many stakeholders and has been extensively analyzed for many decades with advanced computational methods. Textual financial reports and news contain not only the factual information about events, but also explain why they have happened. Exploiting finance and business related textual information in addition to numeric financial information could potentially increase the quality of decision-making. Researchers are searching for effective and computationally fairly simple tools that would be able to handle sophisticated text-related tasks without thorough linguistic preprogramming.

The message, stylistic focus, language and readability of financial reports are good indicators of the perspectives and developments of any company. These indicators can guide companies' decision makers to more efficient actions in the dynamic business environment. Although, financial experts and experienced readers can detect them and make more precise financial decisions, the manual analysis of textual reports require a lot of time, and time is a costly asset in a financial community. Text Mining methods aim to offer opportunities for automatic analyzing and discovering previously unknown patterns in text. Therefore, the less expensive computer-based solutions for mining financial texts for hidden indications of companies' perspectives are needed.

In this paper, we have studied the language and contents of quarterly reports using linguistic and text mining methods. We have compared the results obtained from linguistic analysis of quarterly reports by means of collocational networks and the results obtained from automatic text mining analysis of quarterly report by means of prototype matching clustering. Our objective was to study how well the computer-aided text-mining tool can perceive the content of quarterly reports in comparison with linguistically motivated collocational networks that outline the most frequent and significant words in the texts. The purpose was to see how meaningful the prototype matching clustering is from the perspective of collocational networks linguistic analysis. We performed the study on the quarterly reports from three leading companies in the telecommunications sector, Motorola, Ericsson and Nokia, for the years 2000-2001.

Our results are somewhat controversial. Some of the reports from the companies have as their closest matches the reports with similar collocational networks and some do not have.

Keywords: text mining, annual reports, prototype-matching clustering, collocational networks, collocations

TUCS Laboratory
Data Mining and Knowledge Management Laboratory

1. Introduction

A huge amount of electronic information concerning companies' financial performance is available in organizational databases and on the Internet today. Numeric financial information is important for many stakeholders and has been extensively analyzed for many decades with advanced computational methods. Textual financial reports and news contain not only the factual information about events, but also explain why they have happened. Exploiting finance and business related textual information in addition to numeric financial information should increase the quality of decision-making. Constantly updated text collections have grown so large that there is not enough time to read and analyze them manually. Additionally, the ambiguous structure of texts makes their analysis rather complicated. Researchers are searching for elegant and computationally fairly simple tools that would be able to handle sophisticated text-related tasks without thorough linguistic preprogramming.

The message, stylistic focus, language and readability of financial reports are good indications about the perspectives and developments of any company. These indications can guide companies' decision makers to more efficient acts on the market. Although, financial experts and experienced readers can detect those indications and make more precise financial decisions, the manual analysis of textual reports requires a lot of time, and time is a costly asset in a financial community. Text Mining methods aim to offer an automatic way for analyzing and discovering previously unknown patterns in text Hearst (1999). Therefore, less expensive computer-based solutions for mining financial texts for hidden indications of companies' perspectives are needed.

The most typical company report is without doubt the annual report, which has received a certain amount of attention from linguists and financial specialists. Annual reports, while being important documents to stockholders and financial communities are controversial. They generate disagreement regarding audience, objectives and credibility Thomas (1997). As a genre, annual reports resemble quarterly reports closely. The same writers produce quarterly and annual reports for the same readers within the same community. The reports have a similar structure, conventions, basic functions and communicative purposes but the time spans are different. The study of the linguistic contents of quarterly reports has nevertheless been overlooked in favour of the study of the language of annual reports. In the short-term perspective quarterly reports are important means for companies in appraising past performance and projecting future opportunities to the readers, who primarily consist of investors and analysts. The beginning of every report, known as the manager's/president letter/message to stockholders, contains management's strategy, summary of the financial performance for the year and an attempt to put in perspective the success or failure of the various initiatives of the company Thomas (1997).

In this paper, we study the language and contents of quarterly reports using linguistic and text mining methods. We compare the results obtained from linguistic analysis of quarterly reports by means of collocational networks and the results obtained from automatic text mining by means of prototype matching clustering. Our objective is study how well a computer-aided text-mining tool can perceive the content of quarterly reports in comparison with linguistically motivated collocational networks, which outline the most frequent and significant words in the texts. The purpose is to investigate how meaningful the prototype matching clustering is from the perspective of

collocational networks linguistic analysis. We perform the study on the quarterly reports from three leading companies in the telecommunications sector, Motorola, Ericsson and Nokia, for the years 2000-2001.

The rest of the paper is organized as follows. We start our explanation by giving a short overview of studies relating to analyzing the language of annual reports, as the closest alternative to quarterly reports. Then, we provide a description of the prototype-matching method that we have used for automatic text mining analysis. Next, we introduce collocational networks as a method we used for linguistic analysis. We relate our results from text mining and linguistic analysis by reviewing an example using both methods for analysis of telecommunications companies' quarterly reports. We analyze the results achieved by the two methods and compare them to each other. We conclude with some suggestions for further research.

2. Related Studies

Since the language of quarterly reports has not been studied, our literature review is based on a broad body of literature on the language of annual reports, conducted both within linguistics and business communication studies. A common feature for the studies mentioned here is that they have only concentrated on one part of the reports, the manager's/president letter/message to the shareholders. Our study is different in this respect, as we focus on the whole body of the reports.

Thomas (1997) concentrated on transitivity, thematic structure, context, cohesion and condensation in the language used in the reports. Thomas studied the annual reports of a machine tool manufacturer during a period, which began with prosperity and ended with severe losses. During the time frame of the analysis, the structure of the language used in the reports had changed. According to Thomas' study, an increase in the use of passive constructions can be seen as the profits decrease. There is also an increase in verbs that present the actor (i.e. the company) as "being" rather than as "doing". This indicates that management is trying to present itself as a victim of unfortunate circumstances. This creates an impression of objectivity for the reader, as if the management was presenting plain facts on recent events. On the other hand, when the company was making more profit, it presented itself as aggressive and forward moving through the use of active voice and verbs with both an actor and a goal. A close look at the language structure in the letters to stockholders made by Thomas (1997) showed that the structure of the financial reports might reveal some things that the company may not wish to announce directly to its outside audience. Another conclusion of this study was the confirmation of the Pollyanna Hypothesis[1].

Kendal (1993) introduced the concept of drama when she noticed a similar opposition between the actions of the company and circumstances created by nonhuman agents. Kendall has classified the words and phrases describing actors and objects in the drama into two groups, *God terms* and *Devil terms*. Some examples of god terms are growth, increased sales and competitive position. These words represent concepts that

[1] By studying negative and positive words in annual reports, Hildebrandt and Snyder (1981) induced Pollyanna Hypothesis (Hildebrandt H.H. and Snyder R., (1981), The Pollyanna hypothesis in business in business writing:. Initial results, suggestions for research. The Journal of Business Communication, **18** (1), 5-15). It states that regardless of the financial state of the company, the language in the annual letters will be predominantly positive.

are unquestionably good in the eyes of the company. Devil terms, on the other hand, are terms like losses, decline in sales and regulations.

Other studies have been made with a focus on the relationship between the readability of the annual reports and the financial performance of a company (Subramanian et al. 1993). The annual reports of the companies that performed well were easier to read than those that originated from companies that did not perform well. Studies have also shown that writers of annual reports see the message they put in the report as their personal representation (Winsor 1993). The annual reports are not only the best possible description of a company, but are also a description of a company's managerial priorities. Thus, the communication strategies hidden in annual reports differ in terms of the subjects emphasized when the company's performance worsens (Kohut and Segars 1992). After performing computer-aided content analysis of more than four hundred president's letter to shareholders and examining empirical linkages between themes in annual reports and companies' performances, Osborn et al. (2001) conclude that the text in annual reports reflects the strategic thinking of the management of a company.

Attempts to semi-automatically analyze a company's performance by examining quantitative and qualitative parts from annual reports have been done by (Back et al. (2001) and Kloptchenko et al. (2002). Back et al. (2001) indicated that there are differences in qualitative and quantitative data clustering results due to a slight tendency to exaggerate the performance in the text. Kloptchenko et al. (2002) attempted to explain this tendency using quantitative analysis by means of self-organizing maps for financial ratio clustering, and qualitative analysis by means of the prototype-matching for quarterly report text clustering. In both studies the researchers noticed that the combination of two mining techniques for two different types of data describing the same phenomena could bring additional knowledge to a decision maker. While annual/quarterly reports explicitly state information about a company's past performance, they also contain some indications of future performance, i.e. the tables with financial numbers indicate how well a company has performed, while the linguistic structure and written style of the text may tell what a company intends to do. The study has shown that the sophisticated semi-automatic analysis of the style and content of the financial reports help to reveal insiders' moods and anticipations about the future performance of their company.

3 Methodology

Our methodology section builds on two different methods with the intention of performing two types of text analysis. We use the prototype-matching method proposed by Visa et al. (2000) for computer-aided text mining, and collocational networks proposed by Magnusson (2002) for linguistic analysis.

3.1 Text Mining with Prototype-matching

The prototype is a document or a part of it, which is of a particular interest to a particular user. This prototype is matched with an existing text collection in order to

obtain a cluster of semantically similar documents. The methodology is based on textual collection preprocessing, i.e. word and sentence level processing. We transform every word into a number, taking into account word length in ASCII symbols, and the ASCII value of every character in a word. We create a common word histogram for the entire text collection and choose a suitable Weibull cumulative distribution. Each word after quantization is presented as a bin number and the values of the best-fitted Weibull distribution. We have performed text quantization on the word level, by creating a common word histogram for the entire text collection. The most common words in the text gain a dense resolution in the histogram bins.

We perform similar procedures for converting every word into a bin number on the sentence level, in order to present the whole sentence as a vector. Hereafter, we consider the Fourier transformed encoded sentences as input vectors and choose a cumulative distribution the same way as on the word level. We divide the distribution into logarithmically equal bins, the number of which is equal to the number of all sentences in the text collection. The best-fitted Weibull distribution is found based on the cumulative distribution of the coded sentences and their scalar quantization to equally distributed bins.

In the next phase, we construct individual sentence and word histograms for each document in the collection according to the documents' word and sentence code numbers and the corresponding value of quantization Toivonen et al. (2001). Having sentence and word level histograms allows us to compare documents to each other simply by calculating the Euclidian distances between their histograms. The smallest Euclidian distance between word histograms indicates a common vocabulary of the reports. The smallest Euclidian distance between sentence histograms indicates similarities in written style and/or content of the reports Visa et al. (2001).

3.2 Linguistic Analysis with Collocational Networks

In order to visualise the central concepts and their connections within a quarterly report we used a method originally devised by Williams (1998). In his study, Williams uses the network as a way of exploring the language of science in order to create specialised dictionaries. The main concept in this method is a collocational network. For the purposes of this study, *collocation* was interpreted by Sinclair (1991)simply as "the occurrence of two or more words within a short space of each other in a text". Collocational networks are visual constructions of collocations forming the unique frame of reference for any "word" within a given sub language (Furnas 1987). Collocational networks outline the central concepts in a text, and their textual connections to each other.

It should be noted that the contents of each report were analysed separately. This means that pairs of words which are referred to as collocations in this study are patterns which occur within a single text, and therefore cannot be considered to be typical for English or even business English.

An important factor in this method is the concept of *significant collocation*. Significant collocation takes place when two or more words occur together more frequently than would be expected by coincidence. Following Williams (1998), significant collocation is measured using the *Mutual Information* or *MI score*. The MI score, an information theoretic concept introduced in linguistics by Church and Hanks

(1990), compares the frequency of co-occurrence of node and collocate with the frequency of their occurrence independently of each other. A more thorough description and evaluation of the MI score can be found in Stubbs (1995). It has the same value regardless of which word of a pair is the collocate and which is the node. The MI score is also sensitive to changes in the absolute number of collocates, when the relative proportion of joint occurrences compared to independent occurrences remains the same. In these cases it works "counter-intuitively": decreasing as the absolute number of collocates increases. This means that two words which always occur together get a higher MI score if they occur only once than if they occur more frequently. Because of these drawbacks an alternative to the MI score might be considered for further development of this.

Collocational networks give us an opportunity to examine which concepts are emphasised by the company in a particular report and how these concepts are reflected through the words that constitute the nodes of the network. We can examine which concepts are most frequently linked to each other, by revealing which words regularly appear within a close proximity to each other. This method does not always bring out combinations of words that are perceived by speakers of the language to belong together as phrases or compound words, such as *balance* and *sheet*, unless they occur very frequently in the text. Because the aim of this study is not to find collocations that are typical for business language in general, but central collocations for analysed reports, this limitation is not considered to be a problem.

4 Results

4.1 Text Mining of Quarterly Reports

We encoded every word from the reports, and constructed a common word histogram as the first step of text clustering. Then, we encoded each sentence from the reports and constructed a common sentence histogram, and a unique sentence histogram for every report. In order to obtain the clusters of the closest reports, every quarterly report must be treated as a prototype and matched against the entire report collection. By calculating the Euclidian distance between reports' sentence histograms we can compare all of the quarterly reports in our data collection. For example, for the Ericsson report from 2000, quarter 1, the closest report by content on the sentence level is from Nokia, 2000, quarter 1. The second closest is the report from Nokia, 2000, quarter 3. This means that the Nokia reports from 2000, quarters 1 and 3 and the Ericsson report from 2000, quarter 1 have similarities in sentence construction and word choice, which constitutes the language structure and written style.

The results from text mining of quarterly reports for Nokia, Ericsson and Motorola are presented in Table 1. Each column in Table 1 contains the report-prototype in the header and the four closest matches to it. Quarter names and proper names, e.g. Nokia, Motorola or Ericsson, did not determine the clusters. Thus, the closest matches to Ericsson prototype-reports are not always other reports from Ericsson. The same is true for Nokia and Motorola reports because word choice has a smaller impact on the formed clusters than the sentence construction. Therefore, we

consider the text mining results from sentence level, attempting to justify on what kind of linguistic basis the computer-aided tool chooses the closest matches.

Table 1. The closest Matches to every report in the collection (Sentence level)

Ericsson2000Q1	Ericsson2000Q2	Ericsson2000Q3	Ericsson2000Q4	Ericsson2001Q1	Ericsson2001Q2	Ericsson2001Q3
Nokia2000Q1	Ericsson2000Q3	Ericsson2000Q4	Ericsson2000Q3	Ericsson2001Q2	Nokia2001Q3	Ericsson2001Q1
Nokia2000Q3	Nokia2000Q2	Motorola2001Q3	Motorola2001Q2	Ericsson2001Q3	Ericsson2001Q1	Ericsson2001Q2
Motorola2001Q3	Ericsson2000Q1	Ericsson2000Q2	Motorola2001Q3	Nokia2001Q3	Ericsson2001Q3	Nokia2001Q3
Motorola2001Q2	Ericsson2000Q4	Ericsson2000Q1	Nokia2000Q1	Motorola2001Q3	Nokia2001Q1	Nokia2001Q2

Motorola2000Q2	Motorola2000Q3	Motorola2000Q4	Motorola2001Q1	Motorola2001Q2	Motorola2001Q3
Motorola2001Q3	Ericsson2001Q2	Motorola2001Q3	Motorola2000Q2	Ericsson2000Q4	Motorola2000Q2
Motorola2001Q2	Nokia2000Q2	Nokia2000Q4	Motorola2001Q2	Motorola2001Q3	Nokia2000Q1
Nokia2000Q2	Nokia2000Q1	Nokia2000Q1	Nokia2001Q2	Motorola2000Q2	Nokia2001Q3
Nokia2000Q4	Nokia2001Q3	Ericsson2001Q2	Nokia2001Q3	Ericsson2000Q1	Ericsson2000Q1

Nokia2000Q1	Nokia2000Q2	Nokia2000Q3	Nokia2000Q4	Nokia2001Q1	Nokia2001Q2	Nokia2001Q3
Ericsson2000Q1	Nokia2001Q2	Nokia2001Q3	Nokia2001Q1	Ericsson2000Q1	Nokia2000Q2	Ericsson2001Q2
Motorola2001Q3	Motorola2001Q3	Ericsson2000Q1	Ericsson2000Q1	Nokia2000Q4	Nokia2001Q3	Nokia2000Q3
Nokia2000Q2	Nokia2000Q1	Motorola2001Q2	Motorola2001Q3	Ericsson2001Q2	Motorola2000Q2	Motorola2001Q3
Nokia2000Q3	Motorola2000Q2	Nokia2000Q1	Motorola2000Q2	Nokia2000Q1	Motorola2001Q1	Ericsson2001Q1

Ericsson. In the majority of the cases, the closest matches to Ericsson quarterly reports are the Ericsson and Nokia reports from other periods of time. However, for Ericsson report from quarter 1 (Q1), 2000 the four closest matches are reports from Nokia and Motorola only. The Ericsson report from quarter 3, 2000 has three closest reports from Ericsson from different quarters of year 2000. The most similar report is the one from the next quarter. Furthermore, the report from Ericsson quarter 3, 2000 is the most similar match to the Ericsson report from quarter 4, 2000. Noticeable, Nokia reports disappear from the closest matches for Ericsson, 2000 quarter 3.

Motorola. The majority of the closest matches to Motorola quarterly reports are the Motorola and Nokia reports from different time periods than Motorola prototype-reports. A minority of the closest matches to Motorola quarterly reports are Ericsson reports. The report from Motorola quarter 3, 2001 has fired as the closest one to eleven different reports from Ericsson, Motorola and Nokia, four of those eleven have appeared as the closest ones to Motorola report itself because of the symmetry of Euclidian distance used in the method. The Motorola reports are shorter than Ericsson and Nokia ones.

Nokia. The majority of the closest matches to Nokia quarterly reports are Nokia reports from other time periods. The next closest are the reports from Motorola, and a minority of the closest matches are the reports from Ericsson. The length of Nokia quarterly reports varies significantly, from 2989 words (Nokia Q1, 2000) to 5463 words (Nokia Q4, 2000). Ericsson reports disappear from being among the closest matches to reports from Nokia 2000, quarter 2 and Nokia 2001 quarter 2, but reappears among the closest matches for Nokia report for 2001, quarter 3.

4.2 Collocational Networks for Quarterly Reports

Before the actual analysis could take place we carried out some preliminary measures, i.e. we removed all tables that could easily be separated from the text, and left out some minor tables. This was compensated during the drawing of the networks by leaving out words such as *adjusted, operational* and *non-operational*, which occurred in these tables. During the drawing of the networks a number of other words with little relevance for the report as a whole were left out as well. Words left out were low-content words such as prepositions, articles, conjunctions, words referring to the time span of the report (*quarter, first, second* etc.) as well as words referring to figures or currency.

The initial stage of the analysis consisted of calculating the Mutual Information (MI) score for all words occurring within a span of four words, as recommended by Sinclair (1991) for studying collocations in English. With text sizes of approximately 4000 words, an MI score of 2.00 was found to produce a network of a size suitable for this study. Lowering the score would have brought in words which occur together only occasionally, whereas a higher limit would have produced a network with only the most frequent combinations, leaving out many of the interesting changes which occur among the mid-frequency words.

We approach the collocational networks produced from the quarterly reports from each company in sequence. There are two points of interest in particular where stability or changes can be seen: the structures of the networks and the words they contain. A closer look will be taken at both of these points in the networks created out of each company's quarterly reports.

Ericsson

A brief overview of the collocational networks based on Ericsson's quarterly reports shows that they never exhibit similarity in their architecture. During the period studied here, both the structure and the content of the networks vary considerably. This is also obvious when looking at the text in the reports: during this period the reports undergo several structural changes along with the worsening of Ericsson's financial performance. New headings are introduced and old ones are abandoned or reorganised.

A particularly remarkable change in the networks happens between the third and fourth report for 2000. Structurally, these networks are completely different. There is also a significant difference between the number of lexical items and the lexical items themselves used in the networks. The collocational networks for Ericsson reports from quarter 3 and 4, 2000 are presented in Figure 1 and 2 respectively.

The network for quarter 3, 2000 starts with the most frequent word, *Ericsson*, which is linked to five collocates. One of these collocates, *increased*, is linked to *sales*, which has four other collocates of its own. One of these collocates, *systems*, is linked to *mobile*, which has five more collocates. These linkages mean that the main network for quarter 3, 2000 consists of three parts, connected by collocational pairs. In addition to this, there are several separate collocational pairs and small networks outside the main network.

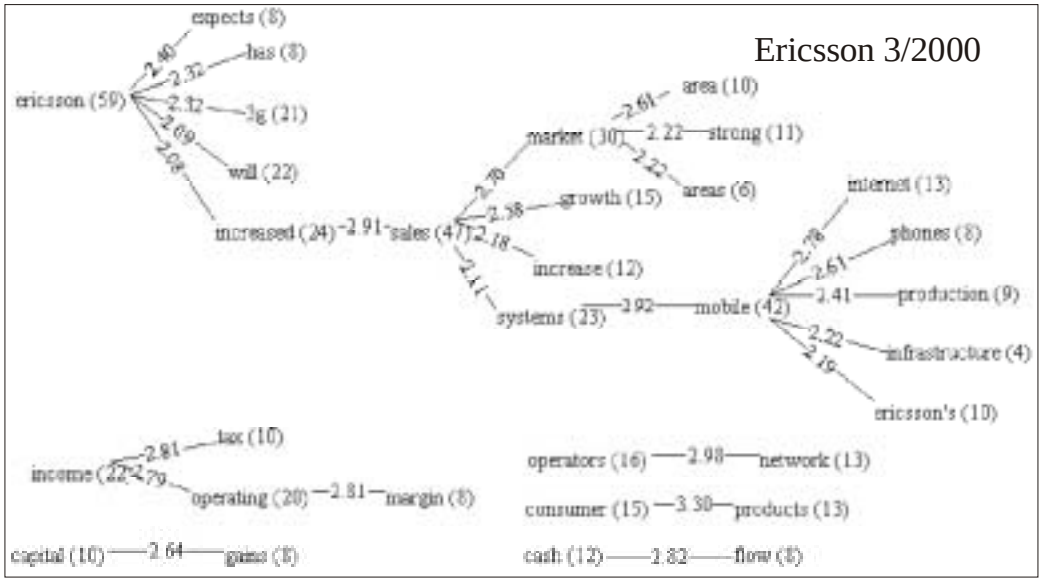


Figure 1. Collocational Network for Ericsson report from quarter 3, 2000

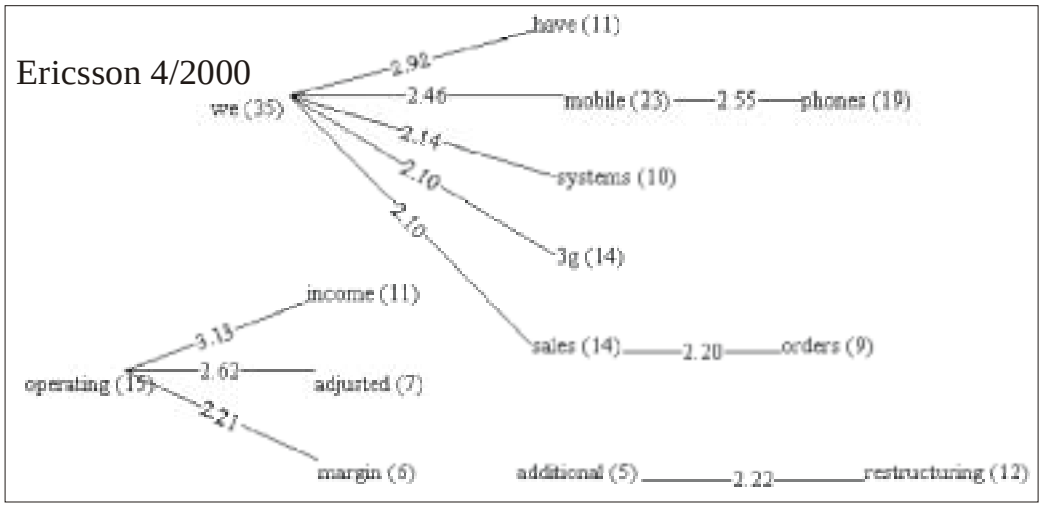


Figure 2. Collocational Network for Ericsson report from quarter 4, 2000

The structure of *network quarter 4, 2000* is very different from the structure of reports from quarter 3, 2000. It consists of a main network attached to the most frequent word, *we*, and a smaller, separate network around *operating*. *We* is a new word in this network, and the most frequent word in network from the report of quarter 3, 2000, while the main word *Ericsson*, has disappeared. Starting from this network, the company now refers to itself using a pronoun instead of the name *Ericsson*. In addition to these two major networks, there is one separate collocational pair, consisting of two new words, *additional* and *restructuring*.

The number of words in the network is much smaller than in the previous network (33 vs. 14), and the structure is much less complex. The most obvious reason for this is

the fact that report quarter 3, 2000 consists of approximately 3600 words, whereas report quarter 4, 2000 consists of approximately 2100 words.

In the following network, quarter 1, 2001, the change continues. This network contains even fewer words than the previous one. Now there is only one word, *expect*, connected to *we*, as opposed to five collocates in the previous network. A new addition is the collocation *efficiency program*, a term bearing obvious negative connotations.
In the last three networks of 2001, more words start to appear and the structures become more complicated. Structurally these networks resemble the networks representing early 2000. Looking at the words they contain, however, they are very different. These networks contain collocations such as *restructuring charges*, *increased borrowing* and *efficiency program*, all of these pointing to a negative development within the company.

Nokia

The collocational networks for Nokia reports from quarter 1 and 2, 2001 are presented in Figure 3 and 4 respectively. The networks are almost identical, containing the name *Nokia* as a central node with links to words referring to the company's business segments such as *Networks* or *Mobile Phones* or general nouns used in business texts such as *sales, market* and *growth*.

However, quite a remarkable change can be seen to take place between the first and the second report for 2001. Structurally, networks quarter 1, 2001 and quarter 2, 2001 look quite similar. They both have just one central word, *Nokia*, around which most of the other words are attached. In addition, there are two collocational pairs outside the main structure. Two words, which appear in quarter 2, 2001, marking the change in the networks, are *decline* and *decreased*. Neither of these words appear in the previous network. In the text of the report for quarter 1, 2001 *decline* does not appear and *decreased* only appears twice, thus making the sudden increase to five and sixteen occurrences respectively is quite an eye-catching. At the same time words bearing positive connotations, such as *growth* and *increased* disappear. The connection between these events is made explicit by the fact that *sales*, a word which is linked to *increased* in the first network, is linked to *decline* in the second.

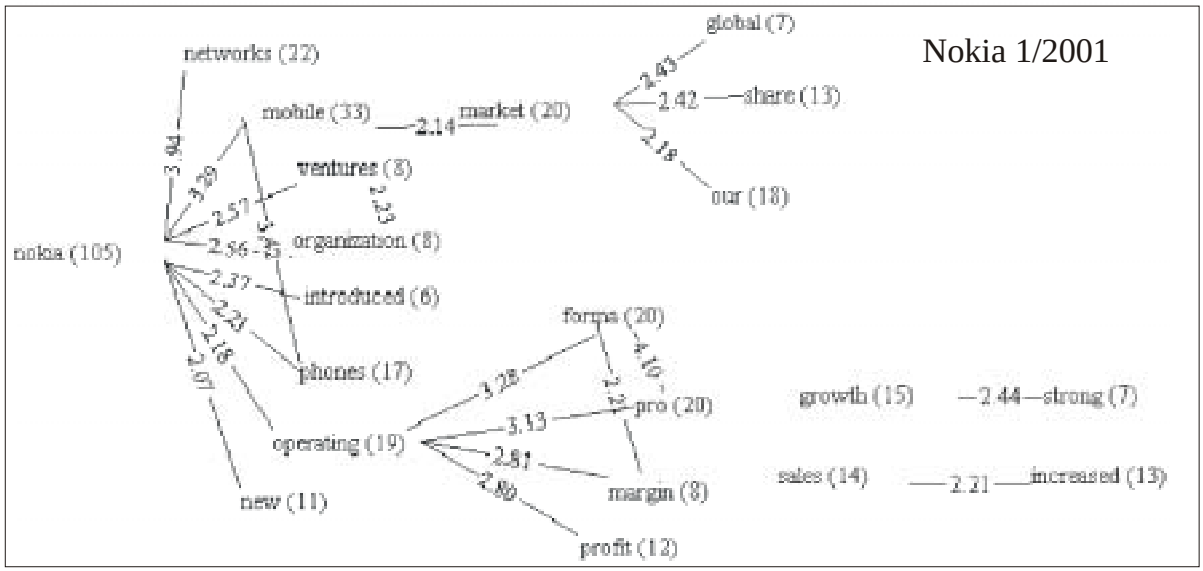


Figure 3. Collocational Network for Nokia report from quarter 1, 2001

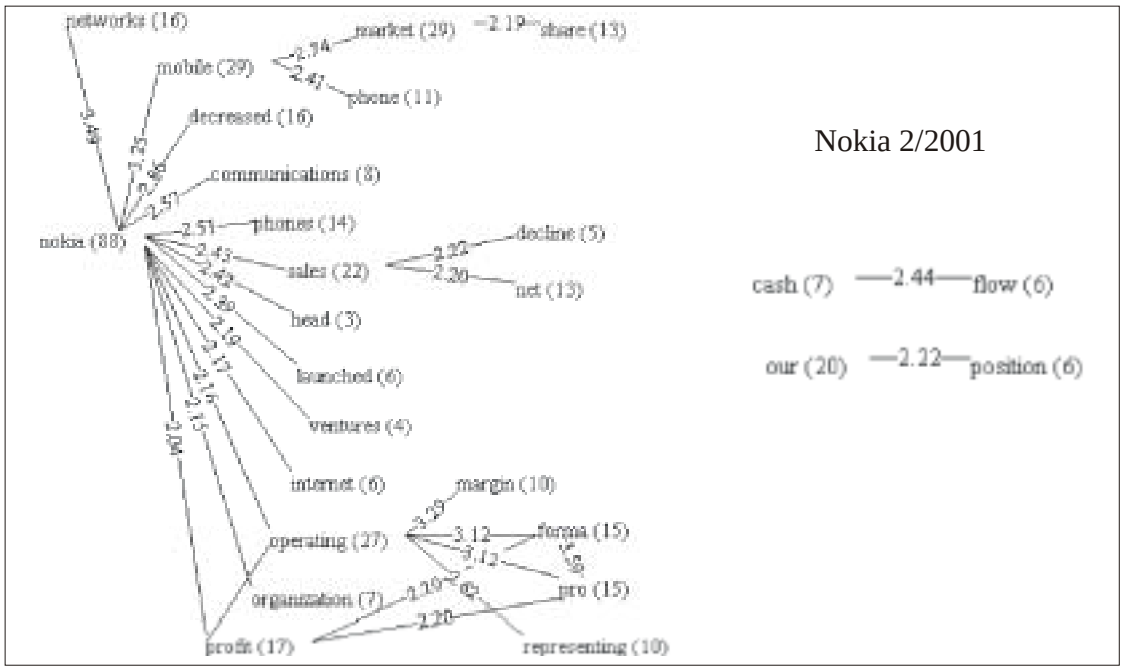


Figure 4. Collocational Network for Nokia report from quarter 2, 2001

Mototrola

The collocational networks for Motorola reports from quarter 4, 2000 and quarter 1, 2001 are presented in Figure 5 and 6. The networks created out of Motorola's reports show less uniformity than the Nokia networks. Still, the networks for the year 2000 resemble each other quite closely. They consist of one main network with the word *sales* as a central node. Linked to it are words like *increased*, *higher*, *orders* and *systems*. Interestingly, the word *lower* appears in these networks.

A budding change can be seen in the first network for 2001. The main network still concentrates around *sales*, but there is also another smaller network around the word *Motorola*, which is the most frequent word in the text and is linked to the collocates *announced* and *new*. It seems as if the company is trying to emphasise the announcements of new innovations. Interestingly, at the same time positive words like *higher* and *increased* have disappeared from the network.

Network 2/2001 looks quite similar. *Motorola* is still the most frequent word, but it is now only linked to *announced*. The word *decline* has also appeared as a collocate to *sales*.

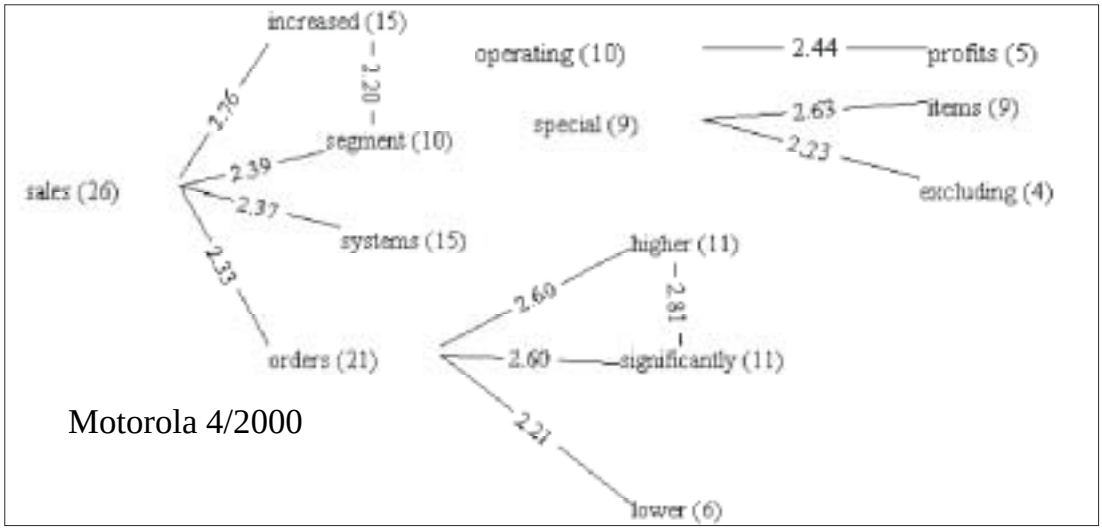


Figure 5. Collocational Network for Motorola report from quarter 4, 2000

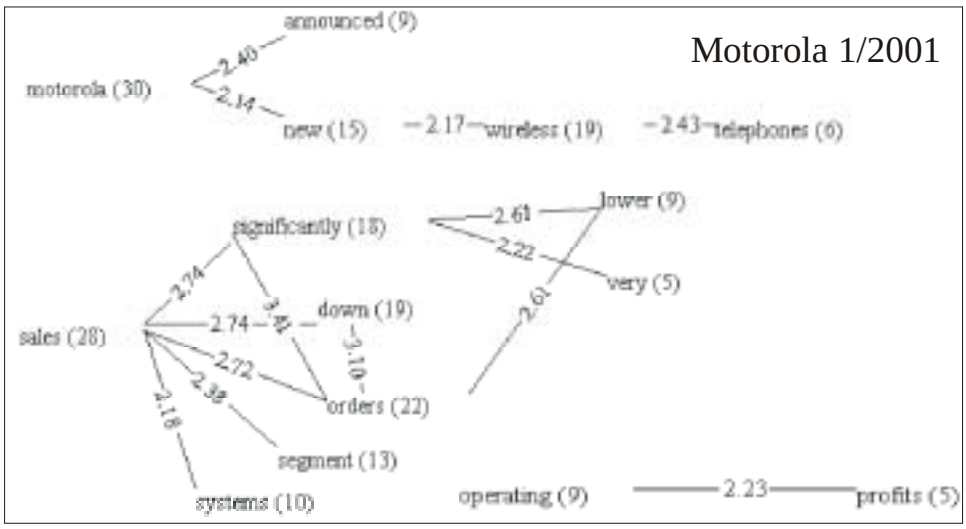


Figure 6. Collocational Network for Motorola report from quarter 1, 2001

In the third network for 2001, illustrated in Figure 7, the changes continue. This network looks very different from the previous ones, as it doesn't contain any structure resembling a network, only pairs of collocations. *Sales*, which was a central node in the previous networks, is now only linked to *segment*. *Motorola* is still linked to *announced*. What makes the contents of this network particularly different from previous networks, is the complete lack of words describing the financial developments of the company, such as *increased*, *decreased*, *higher* and *lower*.

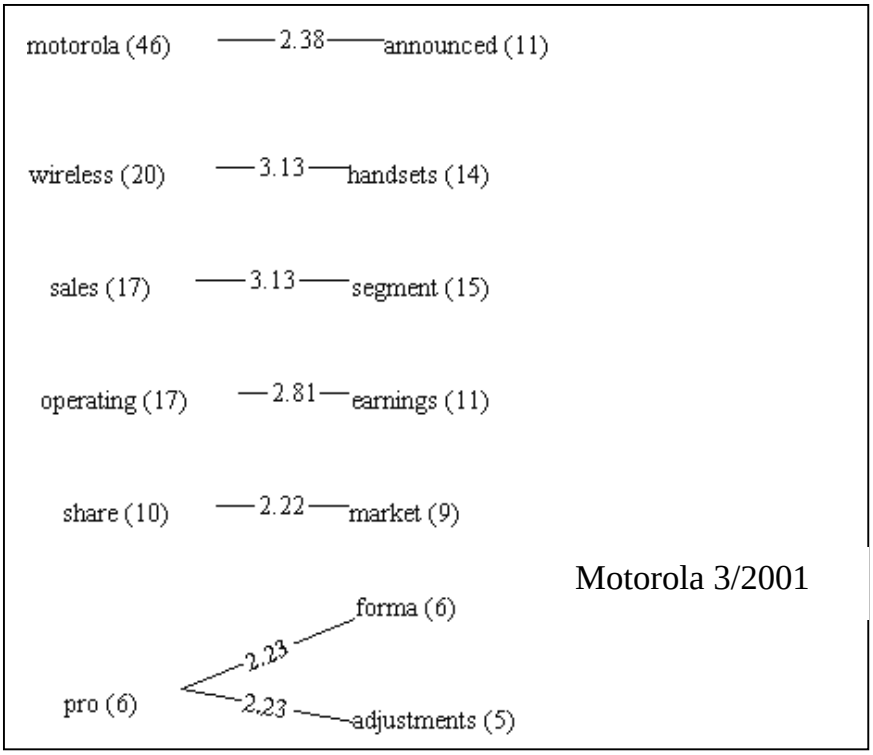


Figure 7.Collocational Network for Motorola report from quarter 3, 2001

4.3 Combining Text Mining Results and Linguistic Analysis

Ericsson

Contradicting to the results of collocational network analysis, Ericsson reports from 2000 quarter 3 and 4 are determined to be the closest matches in text mining analysis.

The collocational networks of the closest matches to Ericsson report from quarter 3, year 2000 are presented in Figure 8. They are the reports from Ericsson 4/2000, Motorola 3/2001, and Ericsson 1,2/2000.

The common parts of the collocational networks for those reports are circled. There is not much similarity between the collocational networks of Ericsson 3/2000 and Ericsson 4/2000 reports as was noticed in the previous section. The common collocates for those reports are *mobile-phone*, *operation*-margin, and *operating-income*. There is no resemblance of the collocational networks between the analyzed report of Ericsson 3/2000 and its other closest match from Motorola 3/2001, because Motorola report has too few links between collocates, that makes the entire network structure weak and links between collocates insignificant for the analysis.

Although the layouts of the collocational networks of the analyzed report and Ericsson reports from quarters 1 and 2/2000 are different, the resemblance is much higher with many common collocates: *sales-grow, sales-increase, sales-increased, sales-market-area, mobile-internet, mobile-infrastructure, income-taxes, income-operating, margin, capital-gains, cash-flow.* The architectures of the collocational networks of Ericsson reports from quarters 1 and 2/2000 are very similar.

The prototype matching method is not able to detect the differences in Ericsson reports from quarter 3 and 4, year 2000, however Nokia disappears from the closest

matches to those reports. The prototype-matching method shows that we have a change in the formation of the closest matches for the report from Ericsson 2000, quarter 2 since other Ericsson reports start to fire among the closest matches. The collocational networks derive this change only by the quarter 4, 2000, by noticing the dissimilarities in collocational network layouts for Ericsson reports from quarter 3 and 4, year 2000.

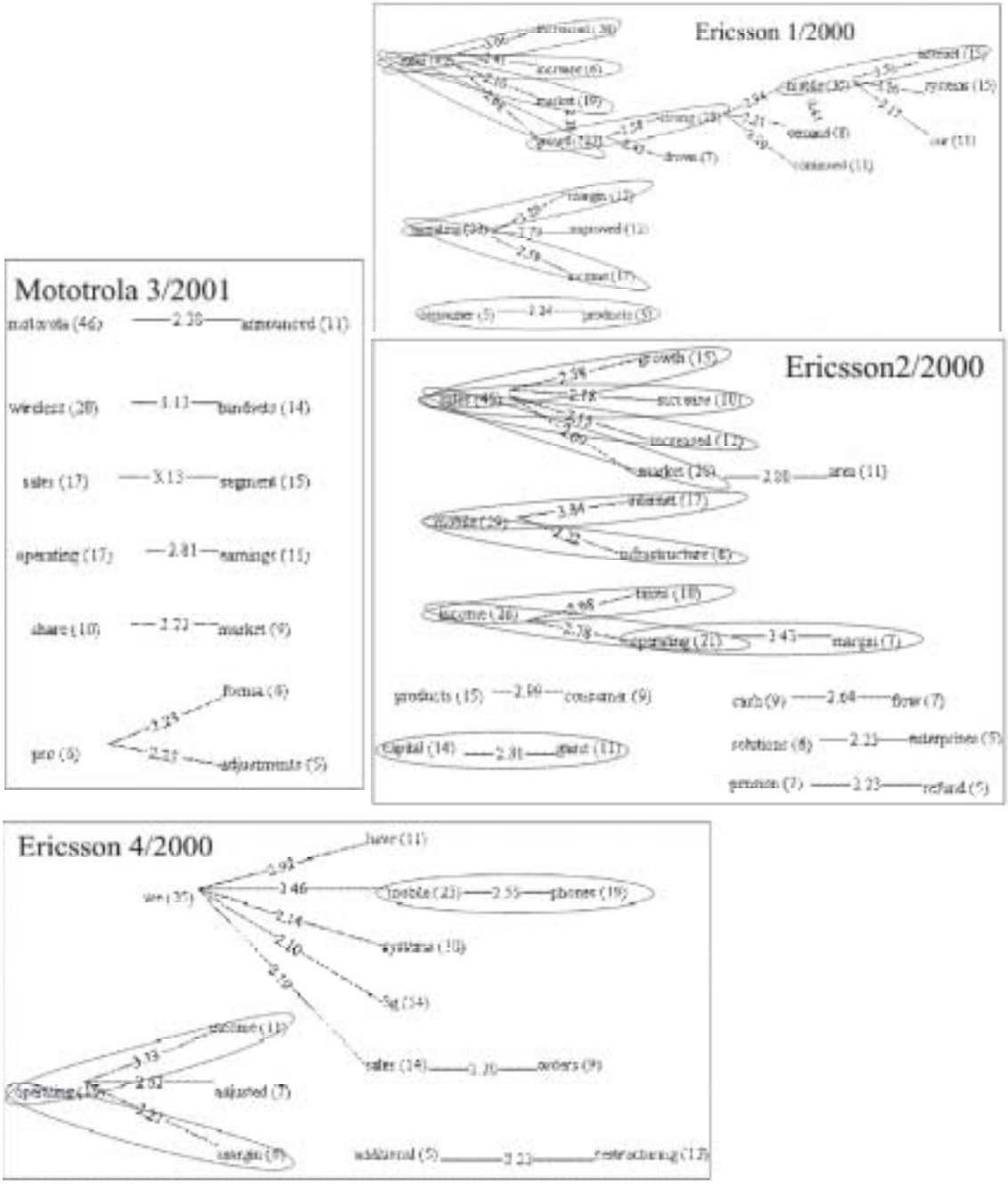


Figure 8. The collocational networks of four closest matches to Ericsson 3/2000 report

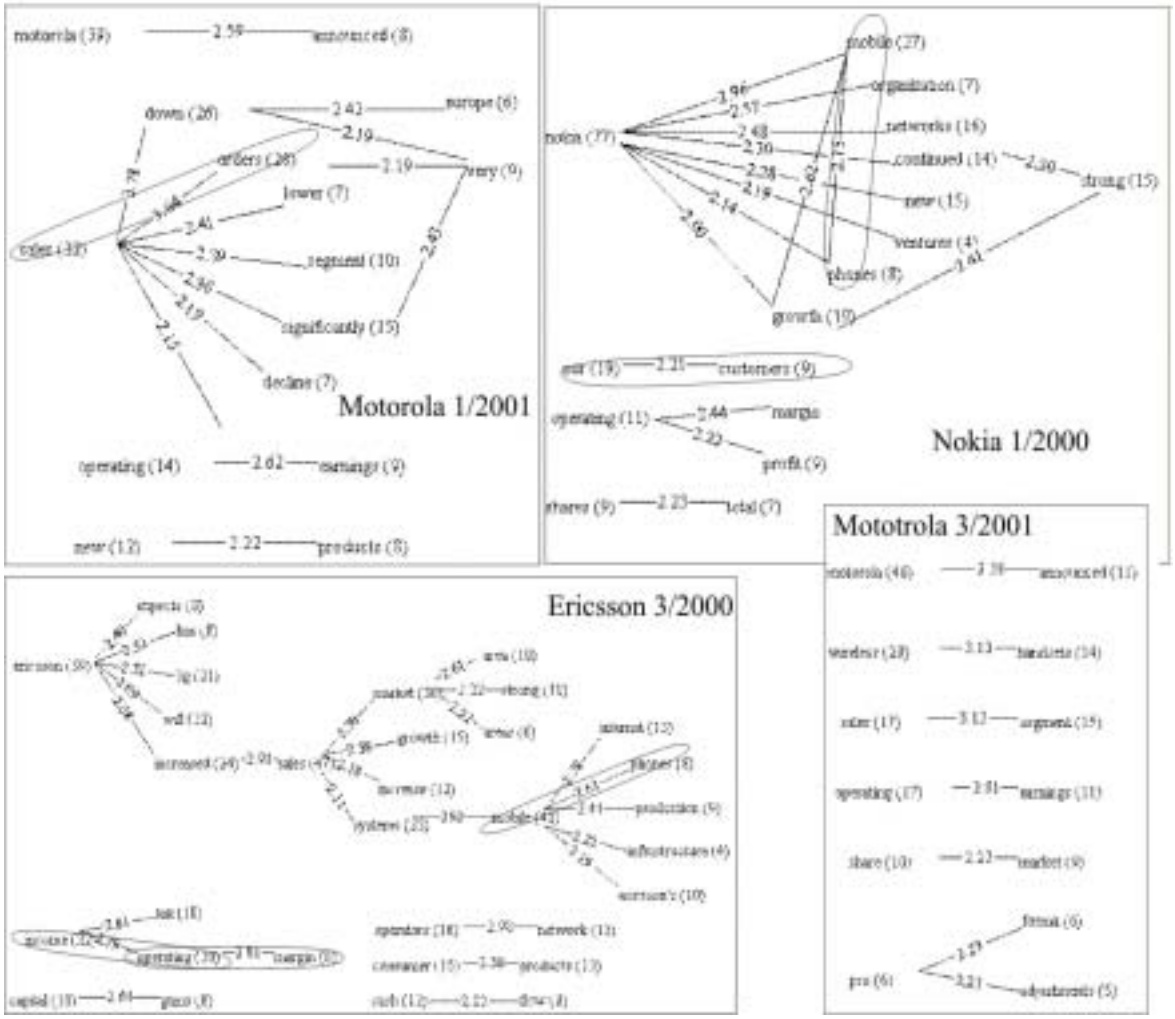


Figure 9. The collocational networks of four closest matches to Ericsson 4/2000 report

There are many more common collocates: *mobile-internet, consumer-products, operating-margin, sales-increase, sales-growth,* and *sales-market.* The common collocates between Ericsson report 3/2000 and collocational networks of Ericsson 4/2000, Motorola 2/2001, Motorola 3/2001, and Nokia 1/2000 reports are circled in Figure 9. Only one (for Motorola 1/2001) or two (Nokia 1/2000 and Ericsson 3/20000) collocates were spotted by the prototype-matching method in the multidimensional structure of quarterly reports. This might mean that either the collocation networks or the prototype matching method omitted some important information coded in the reports, or that different parts of information that was influencing the methods were equally important but simply did not coincide.

The architecture of the Motorola report from 3/2001 has very sparse structure, and thus, was picked out by the prototype-matching method as the closest match to eleven prototype-reports. It means that having only word pairs that were outlined in the collocational networks are not the dominating dimensions upon which the prototype-matching method had performed its clustering.

Nokia

The collocational networks of the closest matches to the collocational network for the Nokia report from quarter 1, 2001 are presented in Figure 10. The common collocates of Ericsson 1/2000, Nokia 4/2000, Ericsson 2/2001, Nokia 1/2000 and the analyzed Nokia report of Nokia1/2001 represented in Figure3, are circled. The resemblance between Ericsson 1/2000 Nokia 1/2001 lies in the following collocates: *sales-increased, growth strong, operating-margin.* There is not much in common between the collocational networks of the analyzed Nokia report and the Ericsson 2/2001 report (only the *operating-margin* collocate). The architectures of the collocational networks of Nokia 1,4/2000 and Nokia 1/2001 are very similar (word Nokia has a central position). Therefore, there are even more common collocates for Nokia 4/2000 and Nokia 1/2001: *mobile-market, Nokia-mobile, Nokia-new, Nokia-networks, Nokia-ventures, Nokia-phones, Nokia-organization, Nokia-introduces, Nokia-operating-profit, market-share.* The resemblance between the analyzed report and the next close match from Nokia 1/2000 is strong: *Nokia-mobile, Nokia- networks, Nokia-new, Nokia-ventures, strong-growth, operating-margin, operating-profit.*

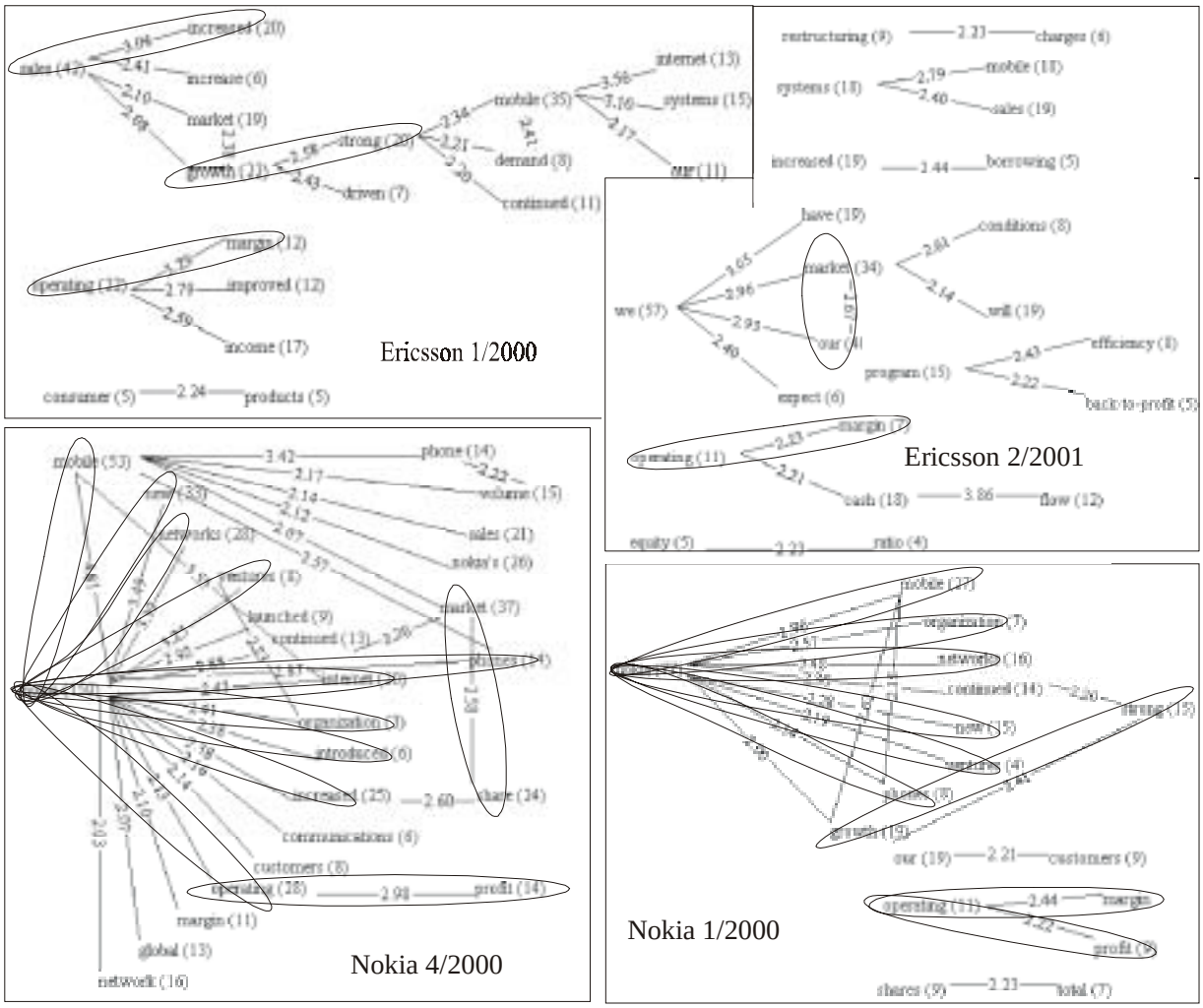


Figure 10. The collocational networks of four closest matches to Nokia 1/2001 report

The collocational networks of the closest matches to the Nokia report from quarter 2, 2001 are illustrated in Figure 11. They are the collocational networks of the Nokia 2/2000, Nokia 3/2001, Motorola 2/2000, and Motorola 1/2001 reports. The common collocates of those reports and the analyzed ones are circled. It is notable that the reports fro Nokia quarter 2 and 3/2001 have a block of similar construction of collocates *operating-margin- forma-pro-representing, profit- forma-pro, operating-representing.* Structurally, the collocational networks of those two closest matches are similar.

Although the collocational networks of all Nokia reports are quite similar, the closest to them, identified by text mining analysis, are not necessarily the reports from Nokia. It appears that the collocational networks of the analyzed Nokia reports bear more resemblance to the collocational networks of Ericsson reports than to Motorola ones, despite the fact that both Motorola and Ericsson reports are among the closest matches.

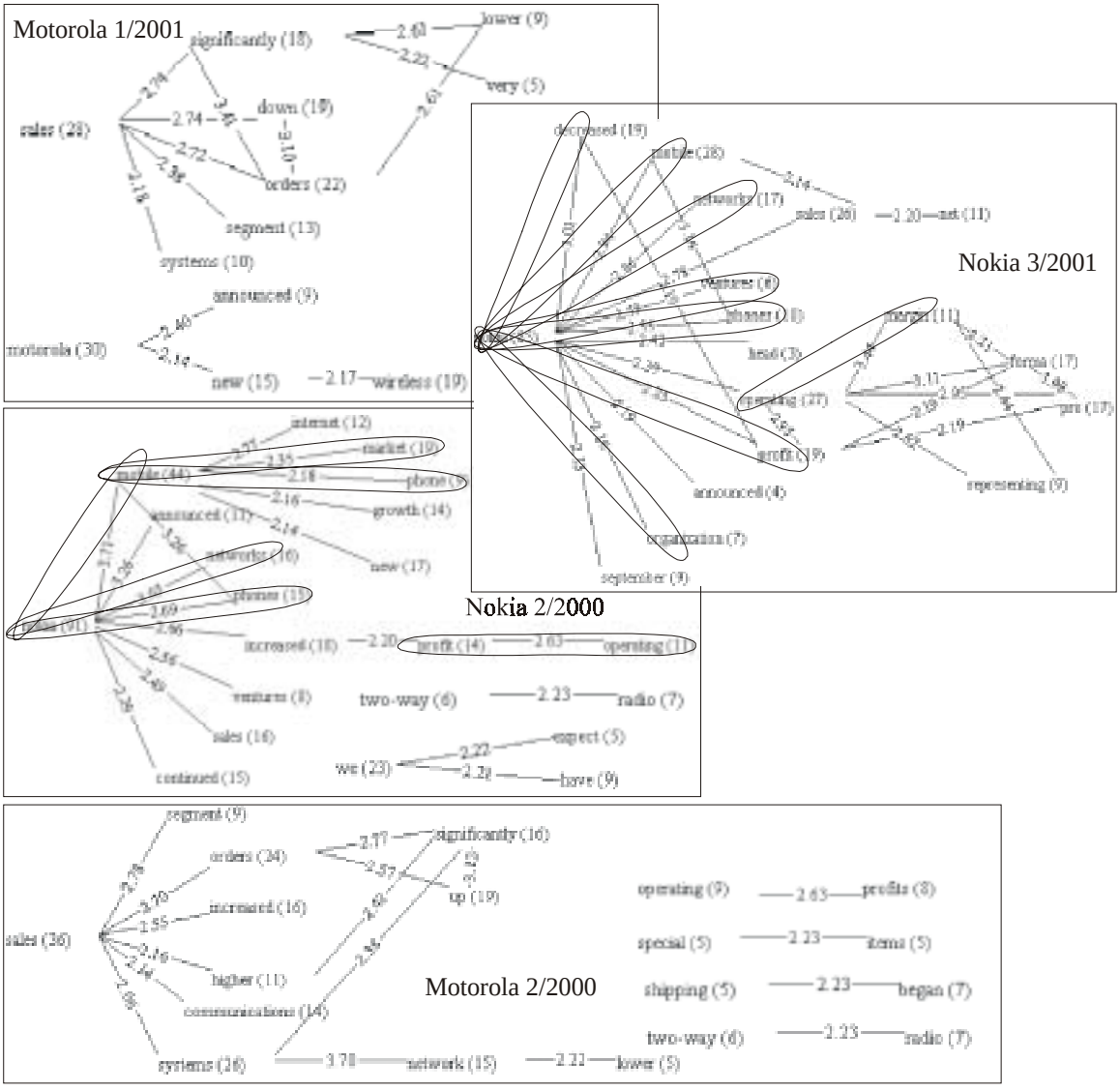


Figure 11. The collocational networks of four closest matches to Nokia 2/2001 report

Motorola.

The collocational networks of the analyzed Motorola reports and their closest matches do not have many common collocates. The common parts between the Motorola 4/2000 and its closest matches: Motorola 3/2001, Nokia 1 and 4/2000 and Ericsson 2/2001 are circled in Figure 12. The resemblance between their corresponding collocational networks is very low. The layouts of the collocational networks of Motorola report 1/2001 and the collocational networks of two of its closest matches from Motorola 2/2000 and Motorola 2/2001 are rather similar (word *sales* has a central position). The common collocates are *sales-segment, sales-systems, operating-profits* and *sales-orders-significantly* (for the Motorola 2/2000 report only), *Motorola-announced* and *significantly-very* (for the Motorola 2/2001 report only). However, there is almost no resemblance between the networks of other closest matches.

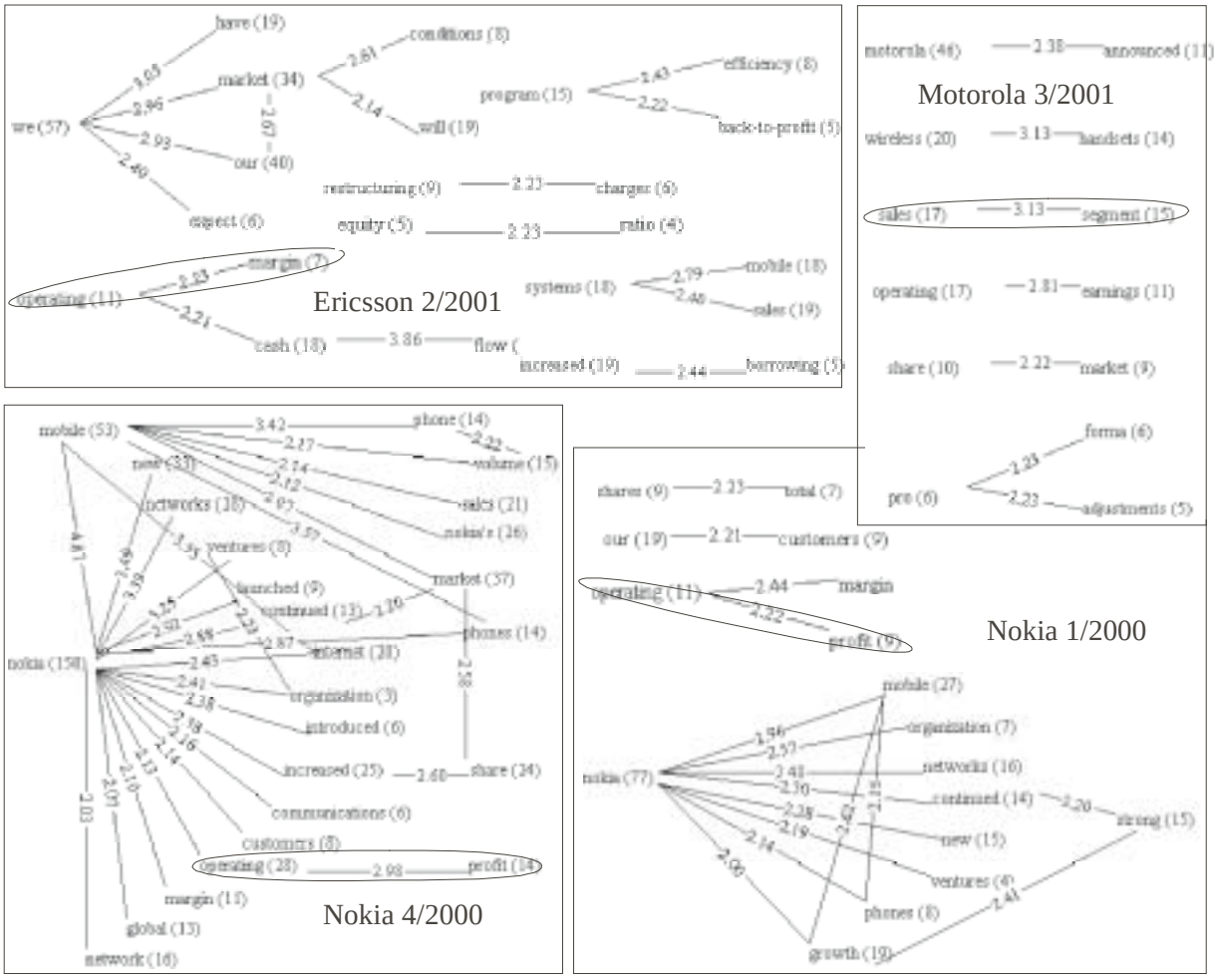


Figure 12. The collocational networks of four closest matches to Motorola 4/2000 report

The common collocates of those reports and the analyzed one are circled and presented in Figure 13. It is notable that Motorola's report from 2001, quarter 3, behaves in an interesting way by firing among four closest matches to eleven quarterly reports in our experiment. This happens because of very many undistinguished collocations with no distinctive network at all. The tone of the Motorola's report is very

neutral and requires further linguistic analysis. The report from the third quarter of year 2001 has no long collocational dependencies and its collocational network looks very different from the rest of the networks. The lack of strong connections between the important terms in this report has resulted in the fact that Motorola 3/2001 has fired as the closest match to eleven analyzed reports (see Table 1).

Collocational networks have illustrated how semantically similar the closest matches reports are to a report-prototype. Because of text multidimensionality establishing adequate similarities between text documents is hardly achievable. Therefore, only two dimensions (*operating* and *margin*) were detected as a similarity reference for Motorola 4/2000, Nokia 4/2000 and Ericsson 2/2001. At the same time, another dimension was spotted by the prototype-matching method for relating Motorola 3/2001 to the analysed Motorola report from 4/2000 – *sales* and *segment*. Although the overall resemblance in words and architecture among the collocational networks of closest matches for Motorola 4/2000 is weak, there is still several semantic similarities between them, upon which the prototype-matching method formed the clusters.

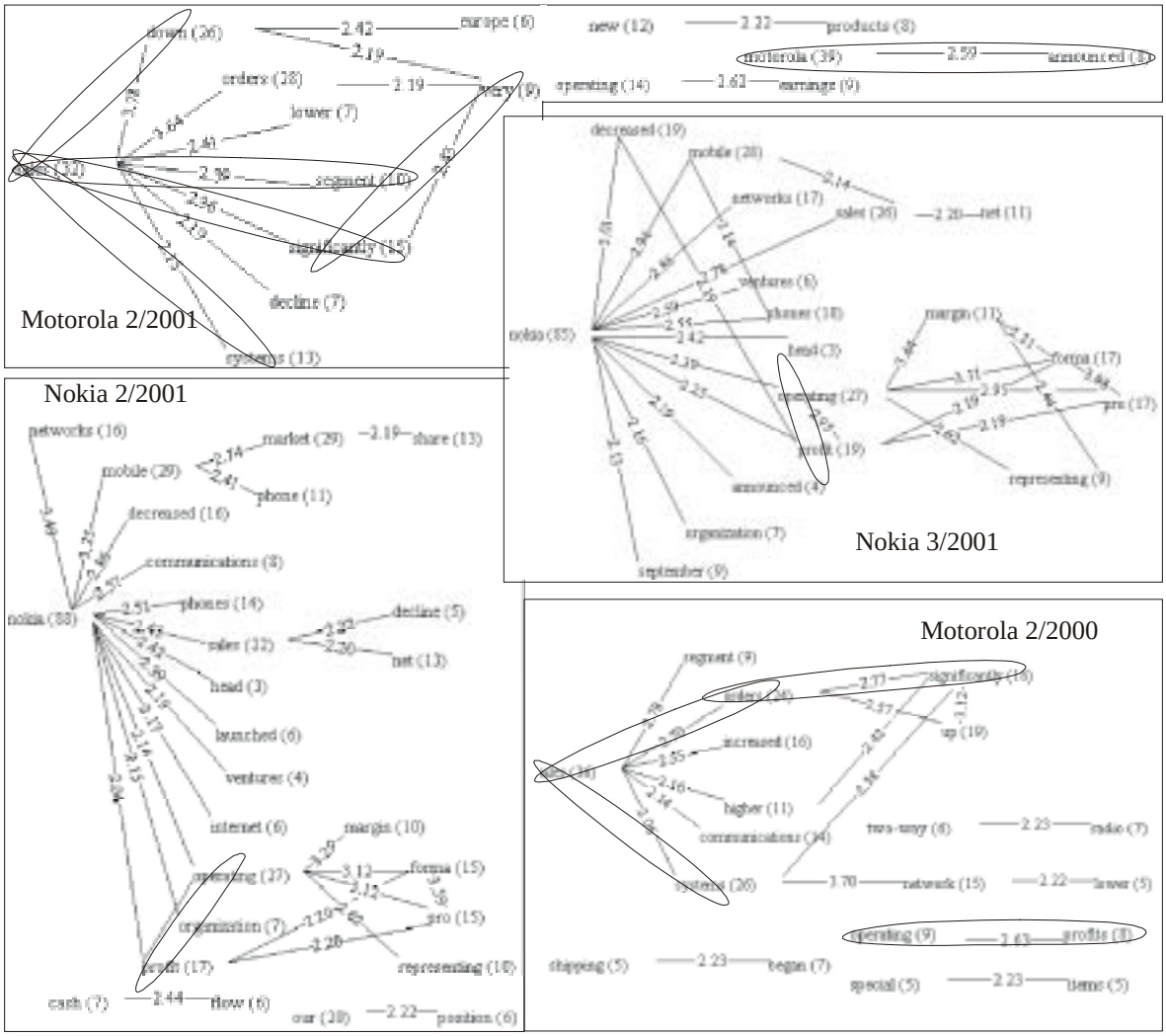


Figure 13. The collocational networks of four closest matches to Motorola 1/2001 report

5. Conclusions and Discussion

Information gathered from all occurring matches from the text mining part of study and the similarity of their collocational networks make it possible to conclude that the computer-aided text mining schema has captured a tendency of changing content in the reports relatively well. Our results from two parts of the analysis have shown to be somewhat controversial. Some of the reports from the companies have as their closest matches reports with similar collocational networks and some do not. Nevertheless, collocational networks and prototype matching text mining aim at presenting and visualizing textual information in a format that can be intuitively recognizable by decision makers. In other words, instead of reading all the reports and trying to compare the companies achievements and determining companies strategies, the decision maker can quickly browse the structure of collocational network or look at what types of reports are similar to the analyzed one.

We realize that the size of our text collection is the biggest limitation for drawing general conclusion. There are some limitations in constructing the collocational networks that affect the ability of the network to outline the central concepts within a report. The comparison results are controversial, because text is a multidimensional data that different readers understand differently. While collocational networks outlined one dimension in text, based on the parameters we had chosen, computer-based analysis took into account several text dimensions.

It will be beneficial to analyze why the closest matches to any chosen report have somewhat different collocational networks. Maybe the closest matches are capturing something more than the structure of the most commonly used terms in the reports. We plan to combine the results from our comparison with the analysis of the actual financial performance of the analyzed companies, i.e. using financial ratios and domain knowledge. The occurrence of the closest matches to any chosen prototype-report possibly contains information on companies future financial performance. In other words, if one knows that Nokia is a financially well performing company, than having closest matches from Nokia to any chosen prototype-report can indicate the semantic similarities that outline good performance. Contrarily, the disappearance of Nokia reports from the list of closest matches, such as for Ericsson report 2000, quarter 3 can imply a decrease or structural change in its financial performance, which was actually detected by the change in the collocation networks of the analyzed Ericsson report and consequent one.

The prototype-matching method compares textual reports by detecting only several similar dimensions from them. While some collocational networks have outlined the same dimensions of closest-matches reports upon which the prototype-matching method performed its clustering, some other collocational networks have outlined different text dimensions. That led to somewhat discordant results in cross-validation, when occasionally collocational networks of the closest matches did not resemble each other.

For future work, a study on the usability of the prototype-matching method and collocation networks that extract the essential key terms from long financial reports by managers can be performed.

6. Acknowledgements

The research was presented at at the XXVI Annual Congress of the European Accounting Association, 2-4 April, 2003, Seville, Spain. We gratefully acknowledge the financial support of TEKES (grant number 47 533). We are grateful to Antti Arppe for his valuable comments and suggestions.

References

Back, B., J. Toivonen, H. Vanharanta and A. Visa (2001). "Comparing numerical data and text information from annual reports using self-organizing maps." International Journal of Accounting Information Systems **2**(4): 249-269.

Church, K. and P. Hanks (1990). "Word association norms, mutual information, and lexicography." Computational Linguistics **16**: 22-29.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T., . (1987). "The Vocabulary Problem in Human-System Communication." Communications of the ACM **30**(11): 964-971.

Hearst, M. (1999). Untangling Text Data Mining. 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), Maryland, USA, ACM Press.

Kendal, J. (1993). "Good and evil in chairmen's "boiler plate": an analysis." Organization Studies **14**: 571-592.

Kloptchenko, A., T. Eklund, B. Back, J. Karlsson, H. Vanharanta and A. Visa (2002). Combining Data and Text Mining Techniques for Analyzing Financial Reports. The 8th Americas Conference on Information Systems, Dallas, USA.

Kohut, G. and A. Segars (1992). "The president's letter to stockholders: An examination of corporate communication strategy." Journal of Business Communcation **29**(1): 7-21.

Osborn, J. D., C. I. Stubbart and A. Ramaprasad (2001). "Strategic Groups and Competitive Enactment: A Study of Dynamic Relationships between Mental Models and Performance." Strategic Management Journal **22**: 435-454.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford, Oxford University Press.

Stubbs, M. (1995). "Collocations and semantic profiles. On the cause of the trouble with quantitative studies." Functions of Language **2**: 23-55.

Subramanian, R., R. Isley and R. Blackwell (1993). "Performance and readability: A comparison of annual reports of profitable and unprofitable corporations." Journal of Business Communication **30**: 50-61.

Thomas, J. (1997). "Discourse in the Marketplace: The Making Meaning of Annual Reports." Journal of Business Communication **34**: 47-66.

Toivonen, J., A. Visa, T. Vesanen, B. Back and H. Vanharanta (2001). Validation of Text Clustering Based on Document Contents. Machine Learning and Data Mining in Pattern Recognition (MLDM 2001), Leipzig, Germany, Springer-Verlag.

Visa, A., J. Toivonen, S. Autio, J. Mäkinen, B. Back and H. Vanharanta (2001). Data Mining of text as a tool in authorship attribution. AeroSense 2001, SPIE 15th Annual International Symposium on Aerospace/Defense Sensing, Simulation and Controls, Orlando, Florida, USA.

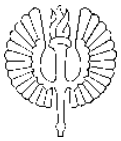
Visa, A., J. Toivonen, B. Back and H. Vanharanta (2000). A New Methodology for Knowledge Retrieval from Text Documents. TOOLMET2000 Symposium - Tool Environments and Development Methods for Intelligent Systems.

Williams, G. C. (1998). "Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles." International Journal of Corpus Linguistics **3**: 151-171.

Winsor, D. (1993). "Owning corporate texts." Journal of Business and Technical Communication **7**(2): 179-195.

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.fi/



University of Turku
• Department of Information Technology
• Department of Mathematics

Åbo Akademi University
• Department of Computer Science
• Institute for Advanced Management Systems Research

Turku School of Economics and Business Administration
• Institute of Information Systems Science