The mood of the (financial) markets: In a corpus of words and of pictures[1]

Khurshid Ahmad, David Cheng, Tugba Taskaya, Saif Ahmad, Lee Gillam, Pensiri Manomaisupat, Hayssam Traboulsi and Andrew Hippisley

Department of Computing, University of Surrey

Abstract

Corpora of texts are used typically to study the structure and function of language. The distribution of various linguistic units, comprising of texts in a corpus are used to make and test hypotheses relevant to different linguistic levels of description. News reports and editorials have been used extensively to populate corpora for studying language, for making dictionaries and for writing grammar books. News reports of financial markets are generally accompanied with time-indexed series of values of shares, currencies and so on, reflecting the change in value over a period of time. A corpus linguistic method for extracting sentiment indicators, e.g. shares going up or a currency falling down, is presented together with a technique for correlating the quantitative time-series of values with a time series of sentiment indicators. The correlation may be used in the analysis of movement of shares, currencies and other financial instruments.

1. Introduction

Financial markets are places where financial instruments are bought and sold. These instruments include shares, currencies, bonds: there are shares traded for individual organisations and traders take options – slang bet – on the aggregate value of key shares e.g. Financial Times Stock Exchange (FTSE). Some of these instruments are traded in millions, others in thousands and yet others in hundreds: the prices of instruments change frequently during single trading or over a longer trading horizon. A set of buying / selling prices of instruments, ordered in time, is usually referred to as a (quantitative) time series. In the financial pages of newspapers, and now on specialised web sites, these time series are either displayed independently or as graphical illustrations within (long) texts.

The buying and selling of instruments in itself causes changes in their value: too many buyers for a share and its value goes up, too many sellers the value goes down. The so-called efficient market hypothesis (EMH) suggests that the (trading in a) financial market is the sole arbiter of the price of an instrument. Despite the preponderance of the EMH, newspapers and financial web sites regularly report the reactions of individuals, acting on their own or on behalf of organisations or governments: some web sites display results of polls of financial experts. The experts report whether they are either 'bearish' – or shy to buy or sell, or 'bullish' – too eager / aggressive to buy or sell, and indeed some of them are neutral. These polls are conducted at regular intervals and the sentiments of experts are displayed as a time series. And, then there are some who 'correlate' the time series comprising the bearish / bullish / neutral voting figures with the time series of financial instruments. The correlation is then used as a cipher for buying or disposing off an instrument.

The sentiment of the market traders – or market sentiment for short – is shaped by, and in turn shapes, the value of a financial instrument usually in the short term and perhaps in the long term as well.

Much as the sentiment of a trader influences others, others also influence him. The view of the others is typically communicated through press statements. One can argue that (financial) news stories may affect the trader's sentiment, or more precisely, his or her attitude towards an instrument. Ergo, positive news stories may persuade people to invest in the market thereby driving the prices of instruments *up*: conversely negative or gloomy stories force prices *down*. Note that the physical prepositions (*up*/*down*) used to describe the position/location of a physical object, are also used to describe the change in value of an abstract financial instrument during a fixed time period. News stories and people's conversation about the financial markets extend these spatial metaphors further by talking in terms of *state change* – one sees a change in the value of an instrument in terms of *rising* or *falling*.

[1] Based on papers presented at two workshops: LREC Event Modelling Workshop (Spain 2002) and Financial News Analysis Workshop, 11th International Terminology and Knowledge Engineering Congress (France 2002).

The use of literary allusions, including bear/bull, vibrant/anaemic, and more colourful slangs, including the phrase *dead cat bounce*, to describe that the upward movement of stock is much like a lifeless object merely moving because of the laws of gravity, shows a creative use of language in the specialist field of financial trading (Ahmad 2002). The sentiment of a trader toward the market may change by reading a news story in that *bullish* stories may cheer him or her up, and *bearish* stories may depress him or her and in turn depress the market (Knowles 1996).

We wish to explore whether it is possible to *extract* the sentiment from a news story through linguistic analysis. It is possible to use the literature on buying/selling in semantic theory (Jackendoff 1991) as a framework for analysing the meaning of the news stories. The literature on natural language processing (Simmons et al 1984) and on knowledge representation suggests that frame semantics has been used to build systems that can, in principle, analyse, extract and disseminate the meaning (intent?) of a specialist news report. Frame semantics has a number of limitations, and a prominent one is the need for a lexicon that is rich and extensive in terms of meaningful data.

What about a purely lexical approach? Recall, Quirk et al's observation that relates frequency of a lexical item to the *acceptability* of that item by an educated-native speaker of the language (1985). Stretching this dictum to financial markets, one can argue that higher frequency of phrases that express positive sentiment suggests that all is well in the market. Similarly, a predominance of negative phrases might suggest that all is not well in the market and that it may have fallen or is about to fall.

We have analysed three year's output of Reuters financial news comprising over 10 millions tokens published during 2000-2002.

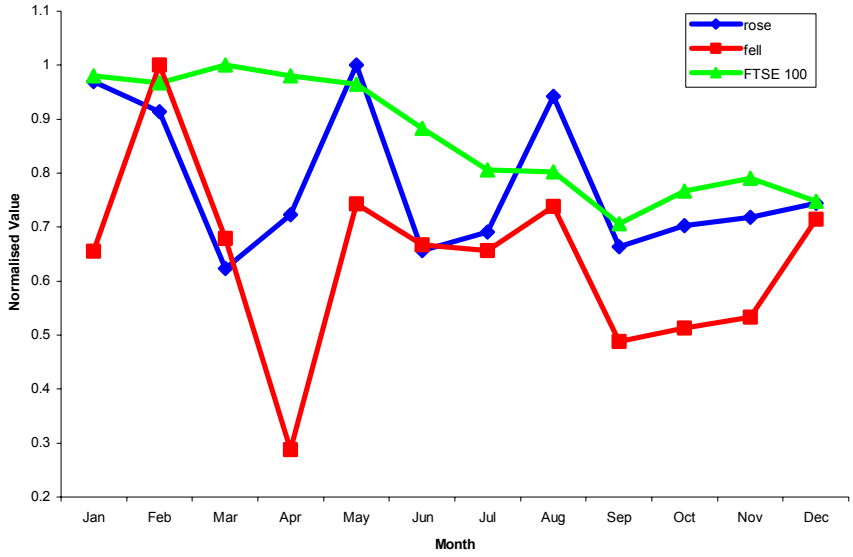


Figure 1: Monthly variation of "rose", "fell" and FTSE 100

Figure 1 shows the variation in the frequency of two verbs *rose* and *fell* over a one-year period (Ahmad et al 2002). Also plotted is the value of FTSE-100 at the close of daily trading during 2002. There appears to be an encouraging numerical correlation amongst the sentiment verbs and FTSE-100.

One further exploration of the above hypothesis that sentiment may correlate with frequency of phrases that may express positive and negative sentiment used to describe changes in the value of an instrument, requires an understanding of the following:

How to analyse texts such that frequent phrases hypothetically related to a sentiment are indeed used in a sentence or in many sentences to express the sentiment? The ambiguity of language does play a significant part in confusing the sense of these phrases, for example, *rose*, as in something *rising*, with

that of a flower or indeed the name of a person. We show how simple cues related to the grammatical categories of phrases in the neighbourhood of the potential sentiment laden word ensure that the word *rise* quite *x* suggests that the value of a financial instrument is rising.

How to use the time series of sentiment word frequency in conjunction with the time series of the values of an instrument either to predict stability or chaos in the market, or to discover the 'turning point' in the value of an instrument? The turning point is the point in time when the value of the instrument stops decreasing and starts to increase or vice versa.

How to organise texts in a diachronic corpus such that these texts may be added or omitted according to the pragmatic attributes of the texts? This is important if the market movements and sentiment analysis are required for a specific instrument (e.g. US$, Euro, BT stocks, FTSE derivative) traded in a certain country or group of countries.

2. Sentiment words and their frequency

2.1 A note on up-and-down phrases

Amongst the many different phrases chosen to describe the changes in the actual or potential values of financial instruments the phrases *up*, *down*, *rise* and *fell* have an intuitive prominence. There are other related phrases and there are synonyms that are also used: *growth*, *slump*, *jump* and *drop* are good examples here. The British National Corpus (BNC) shows the preponderance in the general language of these up-and-down phrases:

Table 1: BNC frequencies obtained from the on-line version of the corpus; for *rose* and *fell* we restricted the query to verbs only (N_{BNC} = *100,106,008*)

Token	f_{BNC}	**Token**	f_{BNC}
up	207709	down	92285
growth	12794	fell	9563
rose	5566	slump	632

The BNC shows that there is more 'positive' sentiment in the corpus than, say, the 'negative'. However a much more detailed analysis is required before any serious conclusion is drawn from these literally raw figures. A similar analysis of financial news texts shows the dominance of the phrases *up* and *down* but with a somewhat different distribution of the other four phrases. Consider a sample from Reuters UK-financial news for the month of November 2002 (comprising 400,000 or so tokens):

Table 2: Frequencies of 'positive' and 'negative' sentiment phrases based on Reuters UK-financial News November 2002. This is an untagged sample and homographic conflicts (e.g. *rose* as *noun* or *verb*) have not been resolved ($N_{Reuters}$=*402089)*

Token	$f_{Reuters}$	**Token**	$f_{Reuters}$
up	1435	down	716
growth	650	fell	391
rose	424	slump	73

The rank order of the three 'positive' sentiment phrases and that of the 'negative' sentiments is preserved when the register changes from the BNC – a corpus which largely comprises general language texts (c. 62% texts are drawn from *fiction*, *leisure*, *world affairs*, *the arts*, *belief and thought*) – to the specialist financial news which wholly comprises, what the BNC compilers would call, *commerce* and *finance* texts (the BNC has just under 8% of such texts in its composition). However, the relative distribution of the different phrases within the 'positive' and 'negative' sentiment categories is substantially different:

Table 3: Relative distribution of "positive" and "negative" sentiment phrases across BNC and Reuters UK-financial News November 2002.

Token	f_{BNC}	$f_{BNC}/N_{BNC}^{Positive}$	$f_{Reuters}$	$f_{Reuters}/N_{Reuters}^{Positive}$
up	207709	91.8	1435	57.2
growth	12794	5.66	650	25.9
rose	5566	2.46	424	16.9
Total	$N_{BNC}^{Positive}$ =226069	100%	$N_{Reuters}^{Positive}$ = 2509	100%

The predominance of *up* and *down* is reduced by about one-third and the more domain-specific *growth* and *rose* increase dramatically when we change the register from general language to special language: from 6% in the BNC to about 26% in Reuters for *growth* and from under 3% to about 17% for *rose*. Similar results can be obtained for negative sentiment words.

Financial journalists typically express *rise* and *fall* in percentage terms: this gives their story a quantitative and objective look and feel perhaps. A concordance of Reuters UK-financial News (November 2002) containing the phrases *rose* shows a varied usage of the pattern:

rose [∅ / by / only/nearly / to] X percent

Some examples of these patterns include:

enterprise shares	rose	2.83	percent	to 584 pence in
volatile mortgage payments ,	rose	by 0.1	percent	on the month to
home loan repayments ,	rose	to 2.3	percent	in the year to
152 pence, logica	rose	over eight	percent	and cmg climbed 6
property companies --	rose	nearly seven	percent	to 1,235
last week house prices	rose	only 1.4	percent	last month - -

A further analysis of the stories published by Retuers, under the UK-financial rubric throughout the calendar year 2002, shows that not all the above patterns are used as frequently and that one pattern *rose X percent* tends to dominate – or, in other words was used preferentially by the journalists in the year 2002.

Table 4: Variations of sentiment *rose* across year 2002, where f_x denotes the raw frequency of the phrases

Patterns	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
f_{rose}	417	369	245	263	376	245	427	351	357	342	424	231
X %	58.0%	52.0%	49.4%	58.2%	57.4%	53.1%	50.1%	51.3%	44.8%	59.4%	50.7%	63.6%
by X %	2.2%	7.0%	0.8%	1.1%	1.9%	1.6%	1.9%	2.8%	3.6%	2.9%	3.8%	3.5%
to X %		4.1%	1.6%	1.5%	1.6%			0.3%	0.3%	2.3%	2.8%	
over X %	0.5%		0.8%		0.8%	0.4%	1.4%		0.6%	0.6%	1.4%	0.4%
Nearly X %		0.3%	0.4%		0.3%	0.4%	0.5%	0.3%		0.6%	0.7%	
by over X %										0.0%	0.2%	
only X %							0.2%	0.6%	0.3%	0.0%	0.7%	
$f_{rose\ [phrase]\ \%}$	284	238	143	174	238	149	254	217	190	236	256	175
Proportion of rose [phase] %	68.1%	64.5%	58.4%	66.2%	63.3%	60.8%	59.5%	61.8%	53.2%	69.0%	60.4%	75.8%

For example, in January 2002, of all the patterns ***rose [phrase] X percent***, over 58% were *rose X percent*, followed by 2.2% comprising *rose by X percent* and just under 1% comprising *rose over X percent*. During the year 2002, the frequency of pattern *rose by X %* was above 50% for all months except March and September where it was 49.4 % and 44.8 % respectively (the average value was 54%). Table 4 shows the monthly variation in the various patterns comprising the verb *rose*.

The *rise*/*fall* and *up*/*down* metaphor may be expressed through other related words and through synonyms. Textbooks on 'report writing' suggest that without loosing accuracy, one may use related words/synonyms instead of repeating the same words. Indeed, *slump* instead of *fall*, and *jump*/*climb* instead of *rise* appear to be good candidates for not only improving writing style but also may make the news report more sensational. We looked into *Roget's Thesaurus* (1980) and the *WorldNet* online to find related words/synonyms for the rise/fall we have discovered in Tables 2 – 4 thus far. It turns out the frequency of the related words/synonyms is much lower (Table 5).

Table 5: Related words/synonym sets of *rise* and *fall*

Lemma	$f_{Reuters}$	**Related words/ Synonyms**	$f_{Reuters}$
fall, inc. *fell*, *falling*	875	*drop*	155
		slump	69
		strike	53
rise, inc. *rose*, *rising*, *risen*	1004	*jump*	133
		climb	75
		lift	33

3. Sentiments and lexico-grammatical patterns?

The domination of one word form over others with related meanings shows perhaps the precision, which came from a lack of imagination some would argue, in the language of financial journalists. This is good news for people interested in information extraction – a branch of computing dedicated to the extraction of meaning from natural language text. The initial results related to the preponderance of one word form over others (Table 5) together with a preponderant lexico-grammatical pattern in which the word form is found (Table 4), suggests to us that when the word forms *rise* and *fall*, or rather *rose* and *fell*, are used, especially followed by a *number* and 'percentage', then they indicate an *upward* or *downward* movement in the value of some part of the financial market – this being true of *shares*, aggregate share index of one business section (cf. Property companies, *telecommunications*) and *house prices*. This still could be an ambiguity though: the report telling us that 'mortgage defaulters fell by X percent' is bad news except if your are in the debt collection business.

Zellig Harris and his pupil Maurice Gross have been keen to suggest that in certain types of text, one may find *local* grammars in operation: certain phrase structures that occur more frequently in one type of text, or one set of text fragments than in the language as a whole (1991). Harris illustrated this point by citing examples of recursive noun-phrases used in biochemical literature to either refer to complex biochemical compounds or complex biochemical processes. Gross (1993) focussed on how we specify time and date and showed cardinal numbers used to denote time and calendrical expressions (day / month / year, century) embedded in their own local grammar. Barnbrook and Sinclair have used this notion to argue that dictionary definitions are also written in a local grammar (1993). Local grammars and the Hallidayion term 'lexico-grammatical' patterns have a certain resonance, and this we have used to explore the grammatical environment of market-sentiment indicating words.

The sentences comprising *rose* and *fell* embedded in the patterns shown in Table 4 were analysed using a reliable part-of-speech tagger – CLAWS . The local grammar of the two-sentiment indicators *rose*/*fell* comprises these patterns (also see):

$$\text{VVD} \left\{ \begin{array}{c} \text{Ø} \\ \text{AV0} \\ \text{AVP} \\ \text{DT0} \\ \text{PRP} \end{array} \right\} \text{CRD} \quad \text{NN0}$$

$$\text{VVD} \quad \text{AV0} \quad \{ \text{DT0} \} \quad \text{CRD} \quad \text{NN0}$$

$$\text{VVD} \quad \text{PRP} \quad \{ \text{AT0} \} \quad \text{CRD} \quad \text{NN0}$$

Here VVD = verb; CRD = cardinal; NN0 = numeral; DT0 = determiner; AV0 = adverb; AVP = adverb particle; PRP = preposition.

Table 6: Grammatical properties of the lexico-grammatical patterns

Pattern	Grammar
rose X %	VVD CRD NN0
rose by X %	VVD PRP CRD NN0
rose to X %	VVD PRP CRD NN0
rose over X %	VVD AV0 CRD NN0
rose nearly X %	VVD AV0 CRD NN0
rose by over X %	VVD AVP AV0 CRD NN0
rose only X %	VVD AV0 CRD NN0

4. "Virtual Corpora"

The designers of individual corpora have discussed the organisation of the text files within a corpus. The early corpora divided texts into informative/imaginative types (cf. Lancaster-Oslo Bergen and Brown Corpora, c.1960's), the texts were then divided into genre, topic and other pragmatic attributes. The subsequent ones took two different approaches: First, genre-based classification was used by some, where texts were classified into books, magazines, personal correspondence and so on (cf. Collins-COBUILD, c.1970-80's); second, topic-based classification was used by the designers of Lancaster-Longman corpus (c. 1980's) wherein texts were divided into subject topics (science, world-affairs, news and so on). The other pragmatic attributes were included as well. The LOB and Brown corpora were developed for the study of (English) language in general and the Collins-CoBuild and Longman-Lancaster for lexicographical purposes. The more ambitious British National Corpus extended the list of pragmatic attributes, and perhaps made the attributes more explicit. The texts can be selected for analysis through a complex query language or by selecting the texts from a file store. For us, there is a hierarchical structure that drives the design of a given text corpus. The selection of the top-level features drives the selection of the other features – these are "doable" tasks if you know the organisation of the corpus you are using. The user of the computer-based corpus must know how the corpus is structured physically and what is more desirable is the use of 'logical' (attribute-oriented) features of the text.

The organisation of a corpus of news reports suggests that sometimes there would be a need to analyse the corpus diachronically, focusing on a given individual/organisation, financial instrument, and at other times there is a need to conduct a synchronic analysis – for example, the analysis of news about all organisations within a given industry sector or all members of a political party. The analysis maybe required based on a query comprising (a number of) keyword(s) that may be used in the indexation of a set of news reports. The permutations of the pragmatic and lexical attributes of a given text are numerous. It is possible that individual users of a text corpus may like to organise the texts according to their own needs. In order to have a user-configurable corpus, the notion of a *virtual* corpus was introduced (Holmes et al 1994) for analysing technical and scientific texts.

The notion of *virtual corpus* is similar to that of a virtual machine: there is in reality only one corpus, but the users can arrange the text attributes in a hierarchy of their choice based on an actually physically extant set of texts, for the duration of their use. This configurable hierarchy will have to be made available through the agency of a program, within a suite of corpus management programs, for producing this virtual corpus. The notion of virtual corpus introduces a shift from the usual pre-defined and explicit corpus hierarchical approach, in that it allows the definition of virtual hierarchies. Texts can then be retrieved by navigating through an organisation – the virtual hierarchy – specified by the user.

We have designed a corpus for use in the automatic extraction of financial information from newspaper texts. The principal source of this corpus are the Reuters News Agency texts – in NewsML format , which is an extension to XML for enriching news stories, conceived by Reuters, developed and ratified by the International Press and Telecommunications Council (IPTC). Atkins, Clear and Ostler discussed criteria for corpus design (1992). Based on this, and with attributes available in NewsML, news can be organized using the six major (pragmatic) attributes as shown in Table 7 below.

Table 7: Pragmatic attributes used for organising NewsML texts

Publisher	Name; Place of Publication; Source of Publication; Date of Publication; Date of Origination
Availability	Copyright Status; Copyright Duration; Copyright Owner; Usage Restriction
Text	Title/Headline; Dateline; Text Type(book, newspaper, journal, etc), Text Mode (written, spoken?), Text Entry (electronic, transcribed?)
Language	Language Name; Regional Variant
Author	Byline; Reporter Nationality; First Language; Editor
Category	Industry Code; Topic Code

From the above set of pragmatic attributes, a news database management system, *Virtual Corpus Manager (VCM)* was developed at Surrey. VCM can be used to organise texts; to share and retrieve texts; to navigate through (content of) the corpus and to impose integrity and security checks on texts. We have used six different types of constraints. Each constraint allows the user to choose one set of inter-related attributes at a time. For example, the users can choose as many major attributes and for each attribute can choose the sub-attributes. One example in the use of VCM is selecting UK-specific financial texts from Reuters daily news stream for a given year.

5. Visualising the mood of the market

Time series analysis is one of the established and persuasive branches of statistics. This analysis is used extensively for analysing 'a sequence of data indexed by time, often comprising uniformly spaced observations' in science, engineering, economics, commerce, biology and in almost every subject. Financial news reports are usually illustrated with a time series of the instrument that is being reported: the share price of a company at the opening or closing of a day's trading plotted over a financial (calendar) year shows the perception of the traders of the financial *health* (another metaphor) of the company. There are time series, which include both opening/closing and day's highs/lows – the Japanese *candlestick* patterns as they are called in the trade.

There are time series of the geometric mean of major organisations whose shares are traded in the market – FTSE-100 share index is the geometric mean of the share prices of 100 leading organisations at the close of daily trading, and there is FTSE index, which is the geometric mean of all-shares traded in the London Stock Exchange at the close of trading. We similarly have DAX-30 for Germany and CAX-100 for France, and there is Dow-Jones Industrial Average (DJIA) in the USA. Financial traders, having sought opinion from statisticians, tend not to deal with the 'raw' data value, but use other statistical measures related to the value of the instrument(s): typically used indices are that of volatility – a measure based on the standard deviation of closing price from its average value in the past few days/hours/minutes; the moving average; and return value which is the (logarithmic) difference between the value of the instrument at time *t-1* and at time t.

The use of other statistical measures of the quantitative changes in the value of instrument(s) are important for us as we try to attempt to incorporate the use of sentiment indicators – or rather changes in sentiment – in an overall financial analysis framework. One such attempt has involved in helping the traders to correlate the quantitative signal, either in its raw form or the derived forms (*return* and *volatility* measures), with the movement of sentiment indicating phrases. The traders typically use two sophisticated computer systems almost simultaneously during a trading session: one screen dedicated to the value of financial instruments – sometimes resolved at 50 values per minute and the other screen dedicated to news streams supplied by Reuters, Bloomberg and others. A typical trader looks from one to the other and then makes his or her decision.

This, rather simplistic view of financial trading has led to the development of SATISFI – which can simultaneously display, or help to visualise the news, the value of an instrument, and the changes in the frequency of sentiment indicators. Table 8 and Table 9 below show the two most frequent sentiment words used in generating the *positive* and *negative* sentiment time series respectively.

Table 8: Dominant Sentiment words *rose* and *up*

	Relative Frequency (10^{-5})											
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
Rose	87.80	79.81	57.94	64.51	85.39	58.41	63.27	86.09	58.28	62.26	63.67	67.25
Up	117.48	135.81	109.40	88.97	96.87	69.78	99.89	134.09	96.92	94.71	99.98	80.70
Total	***205.28***	***215.62***	***167.34***	***153.48***	***182.26***	***128.19***	***163.16***	***220.18***	***155.2***	***156.97***	***163.65***	***147.95***

Table 9: Dominant Sentiment words *fell* and *down*

	Relative Frequency (10^{-5})											
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
Fell	62.45	88.53	69.28	27.06	68.89	69.78	63.02	75.38	51.53	48.81	51.73	68.40
Down	84.09	100.94	68.88	74.89	61.71	92.51	91.92	87.28	101.22	81.78	67.65	85.31
Total	***146.54***	***189.47***	***138.16***	***101.95***	***130.6***	***162.29***	***154.94***	***162.66***	***152.75***	***130.59***	***119.38***	***153.71***

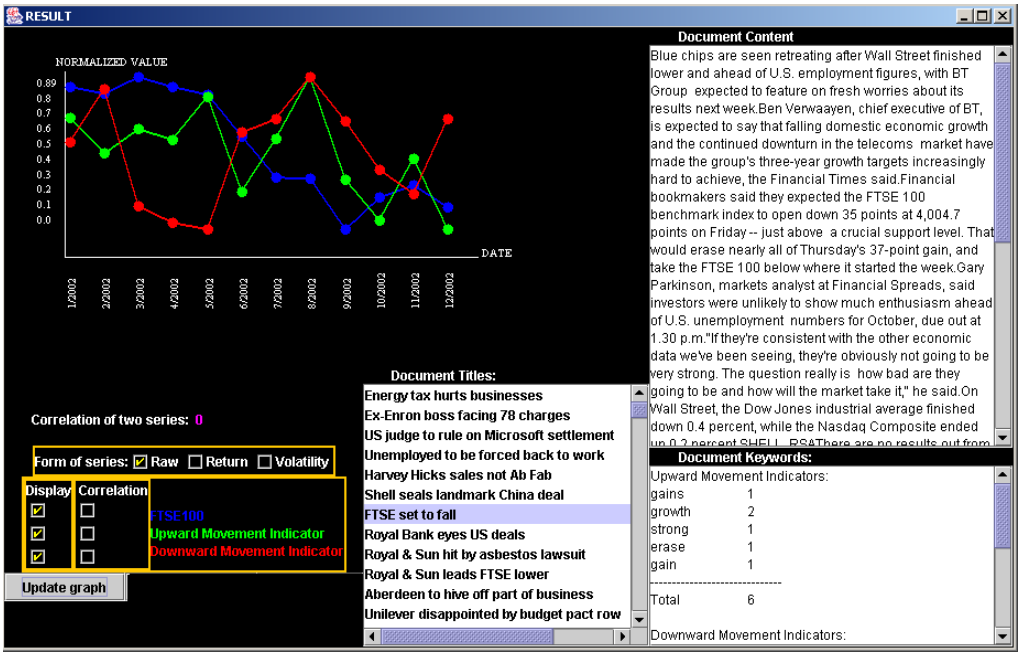


Figure 2: SATISFI prototype shown with one-year FTSE index based on monthly data with upward and downward movement indicator series

SATISFI has four major components that have been fully integrated as shown in figure 2.

i) **Time Series Display**: SATISFI can display three time series at a time. These time series comprise of FTSE-100 close index values, *upward movement indicators* and *downward movement indicators*. As discussed above, upward and downward movement indicators are the quantification of the market sentiment expressed in financial news. Over 70 terms each have been identified for conveying 'good' and 'bad' news. For example upward movement indicators would contain terms like 'up, rise, growth' etc. while downward movement indicators would contain terms like 'down, fall' etc. The movement indicator time series are synthesized by counting these movement indicator terms within the financial news published for a particular day. Each time series is normalised for proper display purposes. SATISFI is capable of displaying the above time series in three forms:
   (1) **Raw** form denotes the original time series.
   (2) **Return** form refers to the logarithmic difference between two consecutive values.
   (3) **Volatility** (historical volatility) is the relative rate at which the time series moves up or down.

ii) **Time Series Correlation:** Correlation is a measure of the degree of linear relationship between two time series. SATISFI provides the user the facility of cross correlating two series in any form (raw, return, volatility). Any series can be shifted forward or backward and cross correlation recalculated to determine whether the market is followed by the news or vice versa.

iii) **Document Display:** This comprises of two parts:
   (1) **Document Titles:** Clicking a dot (date) on any of the time series, displays the corresponding date's news titles.
   (2) **Document Content:** The content of any document title can be viewed by clicking that news title.

iv) **Document Analysis:** Whenever a document title is selected from the news list, the extracted sentiment keywords along with the frequencies are displayed in "Document Keywords" area. Positive sentiment keyword analysis details appear under the title of "Upward Movement Indicators" and negative sentiment keyword analysis details appear under the title of "Downward Movement Indicators".

6. Finding 'meaningful' patterns

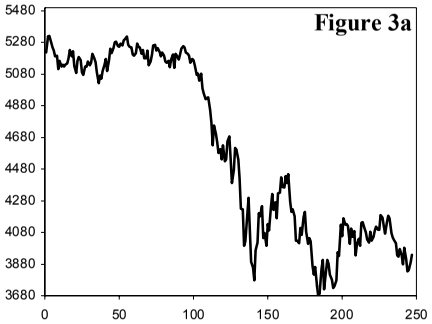
6.1 A case study: Movements in 2002

The year 2002 has seen its ups and downs like many other years and the movements in financial markets worldwide have been in the downward direction. Or, at least, the geometric mean of the value of the shares of the major corporations in North America, Japan, and the European Union, with some exceptions, have reduced substantially. The UK FTSE-100 shows mainly downward movements interspersed by small periods of upward movements, which unfortunately, could not compensate for the previous reduction in the value of the index. In time-series analysis literature one finds techniques that tend to separate the so-called *trends* from *cyclical* movements in the series: the cyclical movements may be due to factors like holidays, trading patterns that may be seasonal and so on (Tino et al 2001). The trends, it is claimed, show a change that is caused because the basic structure of the market has changed. Techniques like *fractal analysis* and the related *chaotic systems methods* help in disentangling the trends from the cycles. Another related and robust technique is the *wavelet analysis* (Rioul et al 1991): a wave comprises oscillations of a number of different frequencies and trends and wavelet analysis suggests ways in which these could be disentangled. We have used wavelet analysis on the FTSE-100 data for 2002 together with the time series of *upward* and *downward* indicating phrases, in Reuters Financial News for the same year. Figure 3a shows the raw figures for the daily trading data for the FTSE-100. Figure 3b shows the long-term trend in the time series (downwards) while figure 3c shows the short-term trend and hence some cyclical behaviour. Note the turning points in the cyclical data (marked by arrows in figure 3c). There is, as noted earlier, a considerable interest in identifying these turning points. The system SATISFI is being extended to generate a textual description of the turning points.

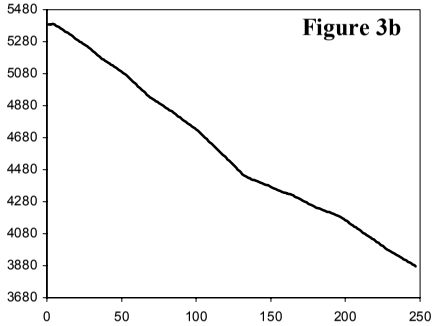
Figure 33a shows a time series of the frequency of a number of upward movement indicators, including *rise, growth* and other less frequent phrases indicating upward movement. There is a corresponding long-term decay in the time-series of upward movement indicators (figure 33b) as found in the raw FTSE-100 data (figure 3b). The cyclical movements are much more pronounced (figure 33c) but a comparison with eye suggests that there is a similar pattern in figure 33c as in figure 3c. The downward movement indicators show that for the first six-months or so of 2002 the frequency of downwards indicators increases rapidly but shows a decay in the later half of 2002 (figure 333b).

The sentiment indicating time-series has to be refined and much more work is required before we may use it to predict the actual mood of the market. However, our approach is perhaps amongst first of the explorations, which investigate how the quantitative movements in a financial market are influenced by

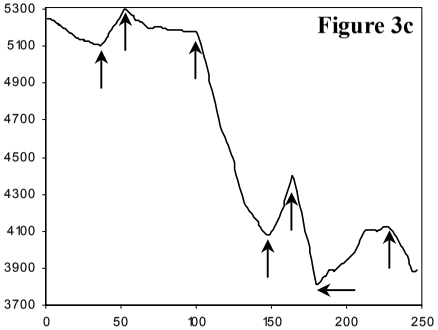
Raw Data: FTSE Daily (2002)



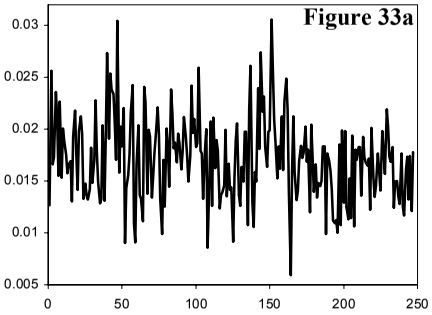
Wav A7: FTSE Daily (2002)



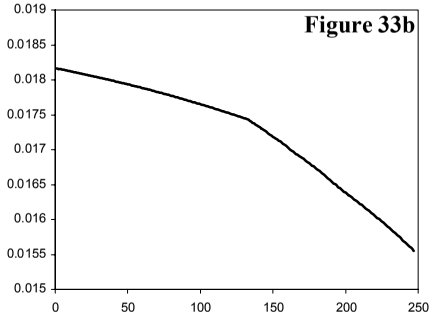
Wav A4: FTSE Daily (2002)



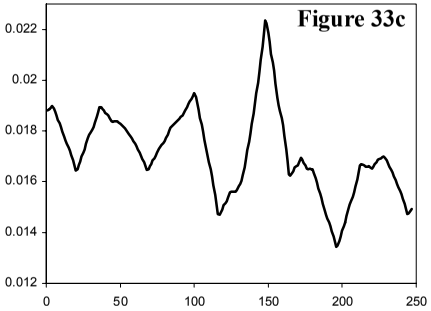
Raw Data: Up Daily (2002)



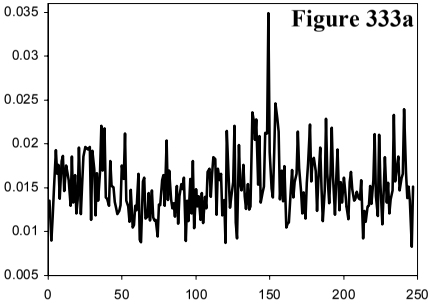
Wav A7 Up Daily (2002)



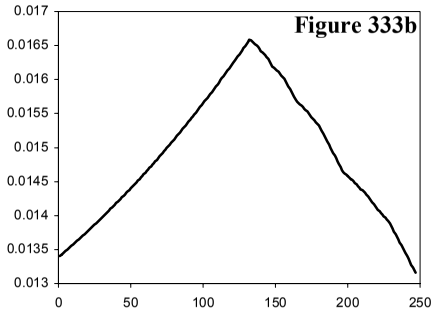
Wav A4 Up Daily (2002)



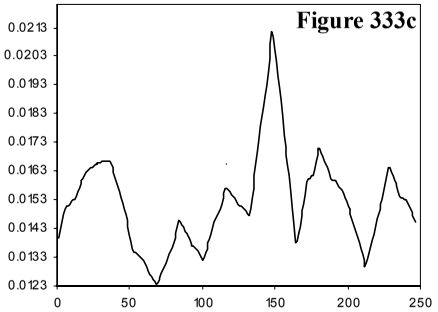
Raw Data: Down Daily (2002)



Wav A7: Down Daily (2002)



Wav A4: Down Daily (2002)



the news stories, some influencing the market and others showing the influence of the market.

7. Afterword

We have attempted to build a system from various linguistic, visual and mathematical components that allows us to explore the behaviour of the traders, and to attempt to model this behaviour. The purpose of the system is to assist the trader by reducing the amount of textual and numeric data that the trader needs to assimilate to form a view of the market. In terms of corpus linguistics, the analysis of qualitative data available in collections of news texts organised by descriptive metadata (through XML and Reuters codes), combined with text processing techniques to determine key patterns, proper-noun analysis to determine key entities and the use of terminology collections for reducing the understanding overhead necessary for the text, and for automatic classification of the text shows early benefits. This work has a direct relevance to the potential for Information Extraction techniques to be adopted across domains, an issue that is being investigated also in a related EPSRC-sponsored project Scene of Crime Information System (SOCIS). The integration of techniques in corpus linguistics with other forms of analysis including mathematical analysis and image analysis provides a supporting environment for future projects that do not focus on a single medium of communication. Experts do not generally rely on the text medium alone; this work provides evidence of the kinds of information fusion that financial experts carry out many, many times each day.

8. Acknowledgements

The work described in this paper has been supported by the EU IST Programme's Generic Information Decision Assistant (GIDA IST 2000-31123).

9. References

Ahmad, K 2002. Events and the causes of events: the use of metaphor in financial texts. In *Proceedings of the workshop at the International Conference on Terminology and Knowledge Engineering*. Nancy, France.

Atkins, S.; Clear, J. and Ostler, N. 1992. Corpus Design Criteria, *Literary and Linguistic Computing* 7(1): 1-16.

Barnbrook, G. and Sinclair, J. 1995. Parsing CoBuild Entries. In Sinclair, J.; Hoelter, M. and Peters, C (Eds.), *The Languages of Definition: The Formalization of Dictionary Definitions for Natural Language Processing*. Luxembourg: Office for Official Publications of the European Communities. pp 13-58.

Gross, M. 1993. Local Grammars and their Representation by Finite Automata. In Hoey, M. (Ed), *Data, Description, Discourse: Papers on the English Language in Honour of John McH Sinclair*. HarperCollins Publishers. pp 26-38.

Harris, Z. 1991. *A Theory of Language and Information: A Mathematical Approach*. Clarendon Press, Oxford.

Holmes-Higgin, P., Abidi S. and Ahmad K. 1994. *'Virtual' Text Corpora and their Management*. In Proc. of Sixth EURALEX International Congress on Lexicography, Amsterdam.

Jackendoff, R. 1991. *Semantic Structures.* Cambridge (USA) & London: The MIT Press.

Knowles, F 1996 Lexicographical Aspects of Health Metaphors in Financial Texts. In Martin Gellerstam et al (Eds.), *Euralex96 Proceedings (Part II)*. Göteborg, Sweden: Göteborg University. pp 789-796.

Maybury, M 1995. Generating Summaries from Event Data. *Information Processing and Management* 31(5): 735-751.

Quirk, R.; Greenbaum, S.; Leech, G. and Svartvok, J. 1985. *A comprehensive Grammar of the English Language*. Longman.

Rioul, O. and Vitterli, M. 1991. Wavelets and Signal Processing, *IEEE Signal Processing Magazine*, pp 14-38.

Robert A. (eds) 1980. *Roget's Thesaurus*. Great Britain, Logman.

Simmons, R. F. 1984. *Computations from the English: A Procedural Logic Approach for Representing and Understanding English Texts*. Englewood Cliffs, NJ: Prentice Hall.

Tino, P., Schittenkopf, C. and Dorffner, G. 2001. Volatility Trading via Temporal Pattern Recognition in Quantized Financial Time Series, *Pattern Analysis and Applications.*

Reuters NewsML Showcase: http://about.reuters.com/newsml/ .
Tagger: http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/trial.html.
System Quirk: http://www.computing.surrey.ac.uk/ai/SystemQ.
WordNet: http://www.cogsci.princeton.edu/~wn/.