Discovering Affect-Laden Requirements to Achieve System Acceptance

Alistair Sutcliffe, Paul Rayson, Christopher N. Bull, and Pete Sawyer
Lancaster University
Lancaster, UK
{a.g.sutcliffe, p.rayson, c.bull, p.sawyer}@lancaster.ac.uk

***Abstract*—Novel envisioned systems face the risk of rejection by their target user community and the requirements engineer must be sensitive to the factors that will determine acceptance or rejection. Conventionally, technology acceptance is determined by perceived usefulness and ease-of-use, but in some domains other factors play an important role. In healthcare systems, particularly, ethical and emotional factors can be crucial. In this paper we describe an approach to requirements discovery that we developed for such systems. We describe how we have applied our approach to a novel system to passively monitor users for signs of cognitive decline consistent with the onset of dementia. A key challenge was eliciting users' reactions to emotionally charged events never before experienced by them at first hand. Our goal was to understand the range of users' emotional responses and their values and motivations, and from these formulate requirements that would maximise the likelihood of acceptance of the system. The problem was heightened by the fact that the key stakeholders were elderly people who represent a poorly studied user constituency. We discuss the elicitation and analysis methodologies used, and our experience with tool support. We conclude by reflecting on the affect issues for RE and for technology acceptance.**

***Index Terms*—Requirements engineering, Affect-laden requirements, Emotional requirements.**

I. Introduction

Novel envisioned systems face the risk of rejection by their target user community, and the requirements engineer must be sensitive to the factors that will determine acceptance or rejection. Conventionally, technology acceptance is thought to be determined by perceived usefulness and ease-of-use [9], but in some domains, other factors play an important role. In healthcare systems, particularly, ethical and emotional factors can be crucial. In this paper we describe an approach to requirements discovery that we have developed for such systems.

In recent years, opportunities have emerged to detect the symptoms of a range of medical conditions by applying computer-based sensing techniques in novel user and social contexts [1]. The work described in this paper took place as part of an investigation into the detection of signs of cognitive decline that may presage the onset of dementia. In addition to the novel technology such a system demands, its development poses interesting requirements engineering (RE) challenges that are, we believe, typical of an emerging class of software applications intended to contribute to health and well-being.

The December 2013 G8 Summit [35] reflected a growing world-wide concern about dementia. In the UK, only 50% of people with dementia ever receive a diagnosis and for those who do, it is often too late for optimal treatment and support. Early diagnosis facilitates interventions which can significantly improve the long-term outcome of (e.g.) Alzheimer's Disease, and treat other disorders such as depression, anxiety and underlying medical conditions, which can lead to reversible memory dysfunction [2].

To increase the number of referrals, it is clear that novel approaches are needed to stimulate greater awareness of cognitive change and the importance of assessment. To this end, the SAMS project (Software Architecture for Mental health Self-management) is investigating the potential of computer sensing for inferring change in cognitive function. However, the efficacy and acceptability of the SAMS approach depend critically on discovery of the user requirements.

The obstacles to understanding these requirements include not only those that are generic to imagined systems for which few contemporary analogues exist, but also a challenging mix of ethical and emotional factors. In addition, the envisioned non-clinical users are predominantly senior citizens, an age group that has been greatly under-represented in requirements studies and for which (e.g.) established techniques such as the Technology Acceptance Model [9] are suspect. In this paper we describe the mix of techniques and technology that we have used to discover requirements for SAMS.

Two contributions are made by our work. The first is that we highlight some of the **requirements challenges** involved in understanding what will make users accept or reject systems they cannot be obliged to use. The second contribution is the **method** for discovery and analysis of requirements for such systems that we have developed and applied in one on-going case study.

The rest of the paper is structured as follows. RE for health and well-being applications, RE for affect, and text mining tools for RE are reviewed in section II. Section III provides a brief description of the project that forms the context of our work. Section IV introduces the method we have developed for the discovery and analysis of a system, the use of which is entirely at the discretion of its target users, who may be

influenced by a mix of concerns. Section V describes in detail how we have applied our method and the results we obtained, both in terms of the requirements knowledge gained and the performance of the support tool used. Section VI discusses the results and section VII concludes the paper.

II. Related Work

In the health informatics literature, requirements are frequently neglected [3], although design rationale in the Goals-Questions-Results method [22] has been used as a shared representation. Jorgensen and Bossen [16] proposed executable use cases coupled to validation via informal animated diagrams for RE in pervasive healthcare. Garde and Knaup [13] argued for a grounded theory approach to RE in healthcare domains to deal with the complexity of the domain and socio-political issues; while participatory design [23] and ethnographic approaches [15] have been applied successfully in healthcare. Cysneiros [8] reviewed a variety of requirements elicitation techniques which could be applied to healthcare, suggesting that technique combination might be more effective. Technique combination (scenarios, prototypes and linguistic corpus analysis) was successfully applied to a healthcare decision support system [33].

A taxonomy of affect-related issues articulated as user-oriented values, motivations and emotions was described by Thew and Sutcliffe [34], with limited explanation of possible implications for values and motivations in the requirements process or for high-level user goals. Ramos and Berry [25] have provided evidence for the impact of emotions on system acceptance. A more detailed taxonomy of social and political RE issues with guidelines for recognising affective reactions among stakeholders was proposed by Ramos et al. [25], [26], who applied their approach in analysing requirements for ERP applications. Callele [6] has applied the concept of emotion in requirements for games applications; however, affect generally has received little attention in RE.

We employ text mining to assist our discovery of affect, where text mining here serves as a convenient umbrella term for any technique classifiable as natural language processing (NLP) or information retrieval (IR). Text mining has attracted significant interest from the RE community and has been applied to a range of requirements problems. Text mining techniques have proven most useful for processing large volumes of text (requirements statements, elicitation transcripts, etc.) where the effort required of a human to process the text is high, and likely to lead to errors due to attention slips and lapses. In such problems, text-mining techniques' inevitably imperfect performance can be most favourably traded off against reductions in effort and human error.

Automatic link generation in requirements tracing [7], [14] is a good example of such a problem. Here, being able to automatically infer *derived_from* relationships by the application (e.g.) TF-IDF is attractive, provided the omission of a minority of genuine trace relationships that the tool will fail to identify can be tolerated [4].

The automatic generation of (e.g. graphical) models by the application of text-mining techniques to textual requirement statements has also generated interest (e.g. [24]). However, the comparative lack of success here illustrates one of the key challenges for text mining in RE. The automatic generation of models requires of NL requirement statements a degree of completeness, precision and absence of tacit information that humans simply don't need [29]. By contrast, humans who have sufficient experience and domain knowledge are able to tolerate a relatively high degree of incompleteness, imprecision and tacit-ness, yet still make a reasonable interpretation of a requirement's intent. There is thus a fundamental mismatch between humans' use of, and need for, economy of expression, and automatic techniques' need for completeness and explicitness.

A consequence of this mismatch is that text that is written for human readers is resistant to the automatic extraction of semantics and pragmatics. This means that the automatic synthesis of requirements from transcripts of elicitation exercises is extremely difficult. Despite this, several text-mining techniques operate successfully at the lexical level to infer *shallow* semantics. Well known examples include Latent Semantic Analysis [10], which has been applied to link generation, and Latent Dirichlet Allocation [5] used in topic modelling.

Corpus linguistics combined with NLP can be used to infer properties of a document by comparison to a large corpus of text whose properties are known *a priori*; this has also found application in RE [27], [30], particularly for abstraction identification [12]. Here, shallow semantics can be inferred from lexical form and context to (e.g.) classify a document focus in terms of semantic categories.

Sentiment analysis [36] has the potential to aid the understanding of affect. However, sentiment analysis focuses on classifying text as expressing positive or negative opinions on a specific topic. This could be useful for (e.g.) crowd-sourcing an assessment of the demand for possible new features. Understanding why and how affect influences stakeholder choice is a different problem, however, and the one we address in this paper.

III. SAMS

SAMS' long-term goal is to help increase the proportion of dementia sufferers receiving an early diagnosis. At its core is a set of passive monitors that collect data as a user interacts routinely with their computer. This data is analysed to infer the users' cognitive health against a set of clinical indicators (CIs) representing (e.g.) memory, motor control, use of language, etc. For example, loss of vocabulary is a common symptom of dementia [18], [20], and this may be discoverable by text mining if (e.g.) the user uses e-mail or social networks to keep in touch with kin. If SAMS accumulates evidence consistent with early dementia, it will issue an alert to the user, suggesting they take a follow-up test such as MoCA [21]. Such tests are claimed to have good diagnostic fidelity and should stimulate the user to visit their family doctor for a full clinical assessment.

There are many technical and clinical hurdles to overcome if SAMS is to work, not the least of which is how to interpret the values ascribed to the CIs from the monitored data. However, even if these challenges can be overcome, SAMS will fail if user acceptance is not achieved. Maximising the prospect of user acceptance was the aim of the requirements elicitation and analysis approach described in the next section. We focused particularly on discovery of requirements arising from:

- users' sensitivity to being monitored and the use to which the data was put, and;
- users' reaction to an alert suggesting they take a follow-up test (and by implication that something was wrong with them).

Both of these are affect-laden. In particular, discovery of requirements arising from the second fall under our classification [32] of *unknown unknowns*, since few users will have experienced anything analogous at first hand.

IV. Methods and Techniques

SAMS requirements discovery was initiated with five workshops that were conducted with a total of 24 participants (14 male, 10 female, age range 60-75, median 66), with a median four participants/session plus two facilitators and one to two moderators from the Alzheimer's Society (AS) or the Dementias and Neurodegenerative Diseases Research Network (DeNDRoN).

All workshops were structured in two sessions lasting approximately 1 hour. In the first session the SAMS system aims, major components and operation were explained followed by presentation of eight PowerPoint storyboards illustrating design options for the alert-feedback user interface, such as choice of media (video, text, computer avatars), content (level of detail, social network) and monitoring (periodic feedback, alert only, explicit tests). The second session focused on discussion of privacy issues in monitoring computer use, data sharing and security, ethical considerations, emotional impact of alert messages, users' motivations and likelihood of taking follow-up tests.

Requirements issues raised in the workshops were explored further in 13 interviews following a similar structured approach of explaining the SAMS system, presenting scenarios to illustrate similar design options with discussion on privacy, security and ethical issues. Questions in the interviews also probed users' reactions to different levels of monitoring (e.g. actions, text) and their perceived trade-off between benefits/motivations versus fears/barriers for adopting the system and taking follow-up action after an alert message. Respondents (4 male, 9 female), ranging from 67 to 89 years old (median 72), were all interviewed in their own homes, apart from three sessions carried out in a community centre.

All workshops and interviews were audio-recorded with the participant's consent. Interview participants were also invited to use an audio recorder for one week after the interview date to record any issues which they subsequently thought might have been included in the interview, and their feelings related to any personal experiences of news stories relevant to the domain, i.e. dementia and Alzheimer's Disease. Two news stories on Alzheimer's Disease (medical progress, personal story of younger patients) taken from the BBC News website were left with the participants if they needed material to prompt reflections.

V. Analysis Results

Analysis of the interviews and workshops are presented in two sections: first a standard requirements analysis treatment of listening to audio recordings and distilling interview notes to produce a list of functional and non-functional requirements. The second section reports more in-depth manual and tool-supported analyses of transcriptions of the interviews and workshops to produce a thematic analysis of requirements issues, stakeholder reactions to designs, emotional feelings about the system, ethical and privacy concerns.

A. Conventional Requirements Analysis

1) Workshop Results

(a) Reaction to Design Options: Opinion was never unanimous on any design option. There was no consensus on choice of media (text/video/avatar), although a majority in all workshops favoured provision of more detail and availability of regular reports (content). In addition, most favoured having an icon serving as a visual cue to remind them that the system was running, with a control that temporarily disabled the monitoring. Use of video was favoured in four workshops where participants suggested that self-help (how to cope) and explanatory videos (dementia mitigation treatments) were important motivators for persuading them to take follow-up action. Active monitoring (e.g. quizzes) was favoured by all, but (e.g. card) games were rejected in three of the five workshops. Participants in all workshops suggested that configuration controls for different design options would be welcome.

(b) Privacy, Ethics and Emotions: All participants expressed concerns over privacy and security arising from monitoring their computer use. Although they were reluctantly willing to share their data with the researchers for analysis, most participants insisted they should have control over their own data. Sharing data with their close kin/friends had to be under their control and the majority would not share information or the alert with their doctor. The majority in all workshops were willing to allow monitoring of their computer use and e-mail text content, suitably anonymised to protect the identities of other parties to conversations. Most participants expected to experience anxiety and fear if they received an alert message, although they all stated that they would take a follow-up test. Contact with a human expert or carer was cited as an important support, with connections to support groups (e.g. the Alzheimer's Society) as additional sources of information to motivate people to take follow-up tests.

2) Interviews

(a) Reaction to Design Options: The interviews produced even less consensus than the workshops for the user interface design requirements. Most respondents (11/13) favoured the

plain text alert message over other media options, although the reminder icon with the disable control (11/13) was a shared requirement with the workshop participants. Active monitoring by a 'cognitive quiz' and a weekly diary was favoured by the majority (11/13) although card games were less popular (8).

(b) Privacy, Ethics and Emotions: The respondents were even more concerned about privacy and security, possibly because three participants had recently experienced phishing attacks on the Internet. However, only two individuals were not willing to have their e-mail content monitored. Opinions on minimal data sharing and the need to maintain control over their own data were similar to the workshop participants'. The majority of the respondents (11/13) expressed anxiety about being monitored, and expected to experience discomfort, fear and worry when they received an alert message, although all these 11 participants stated they would take the follow-up test: 'better to know the bad news' was a common statement. However, ten respondents reported that they could not realistically imagine how they would react in a real-life situation. Five individuals noted that further explanation after the alert message would be vital and all reported that their main motivation for using the system was efficacy: a feeling of being in control by self-management of their health.

3) Requirements Conclusions

Given the diversity of opinion in both the workshops and interviews and inconsistencies between the two modes of requirements capture, prioritising requirements from this analysis was not an easy task. The following requirements list contains issues which supplement the preceding narrative summary of the interviews and workshops. After discussion within the project team, the following conclusions were agreed:

(a) Essential Requirements:

i. Monitoring computer use and e-mail text, but not other applications
ii. Active monitors such as a cognitive/general quiz.
iii. Disable monitoring control always visible
iv. Simple text alert message.

(b) Desirable Requirements:

v. Configuration controls to turn off/on the following options (default off in all cases)
    a. More detailed displays: graphs and chart
    b. Continuous alert icon
vi. Active monitors as diaries, weekly quiz and card games
vii. Video explanation of alert messages and support advice

(c) Additional Requirements- explored but unlikely to be prioritised:

i. Avatar agents for explaining alerts
ii. Chatbot agents to gather text via conversations and explanations
iii. Social media option: closed groups for data sharing and support

(d) Non-functional Requirements:

Privacy and security: controls over any data sharing, encryption and secure transmission of data to university site, encryption on own PC to mitigate hacking attacks, depersonalised data only for wider research sharing.

Maintainability: installing SAMS on user's laptop/PC should not disrupt normal computer use.

Performance: SAMS software should not degrade the performance of the user's machine.

(e) Emotions and Values: Several issues which were categorised as values (see [34]) and emotional requirements [25] appeared to have an important bearing on the requirements and design options.

Trust: in the SAMS system, the universities (system authors), healthcare professionals, follow-up test websites and authors thereof.

Motivations: efficacy, desire for self-control, altruism: participation might help research on dementia.

Emotion: anxiety and fear of negative alert messages, uncertainty over personal reaction.

The non-functional requirements, emotions and values emerged to be both critical to SAMS' acceptance and difficult to get right. These therefore formed the focus of the subsequent thematic analysis.

B. Thematic Analysis

The thematic analysis was performed on the interview recordings and the post-interview recordings, all of which were manually transcribed. The users' text (without the interviewer's contributions) comprised a total of 41,000 words. It was analysed to try to find additional insights into what would help or hinder SAMS acceptance, and to do this in a way that provided quantitative evidence we could use to inform SAMS requirements.

The analysis took two modes:

- a data-driven mode where we mined the text and followed where the data took us; and
- a hypothesis-driven mode where we filtered the text to find evidence of ethical, security and emotional concerns.

In both modes, we performed a manual analysis and a supervised semi-automatic analysis using Wmatrix [27].

Wmatrix provides automated inference of properties of the text, integrated within a framework designed for supervised analysis. It uses techniques from corpus-based natural language processing for the shallow semantic annotation of words and phrases; and a hybrid combination of rule-based and probabilistic approaches to assign a part-of-speech label (e.g. noun, verb) and a semantic field label to each word or phrase in the text. The semantic taxonomy (USAS) [28] used has approximately 230 word-sense classes, each represented by a different tag. It is lexicographically based and derived from McArthur's classification [19] that has been considerably extended and expanded. The semantic tagger aims to identify the coarse-grained contextually correct meaning of a word or phrase. The semantic tagger's accuracy is around 91% on general written and spoken English.

Wmatrix extracts frequency profiles of words, phrases, grammatical and semantic categories and allows the analyst to compare two or more profiles together using the log-likelihood

(LL) metric. This highlights key words or concepts that are unusually frequent in a text relative to a general reference corpus of spoken English.

Fig 1 Wmatrix semantic cloud of the interviews

Figures 1 and 2 show word 'clouds' that are actually formed of the dominant semantic categories in the respondents' interview responses. The larger the font size, the more terms belonging to the corresponding category dominate. In addition, texts can be queried for occurrences of specific semantic categories, such as emotion or knowledge, to see contextual examples of where those concepts occur in the data. It was for this range of capabilities that we selected Wmatrix for the study. Thus a typical Wmatrix *modus operandi* is to list tag frequencies in rank order, or filter on a subset of tags, list the terms to which these tags correspond and then manually investigate instances of these terms in context, using concordances.

1) Data-driven Analysis

The data-driven analysis was used to explore the main themes of the interviews and the post-interview recordings. A manual discourse function analysis was followed by a frequency-profiling analysis.

Interviews and post-interview recordings were analysed first for discourse function at the paragraph conversational exchange level. A subsequent, more detailed analysis focusing on values, motivations and emotions (VME) expressed in short utterances or phrases, was performed as part of the hypothesis-driven analysis, and is described below.

a) Discourse Function Analysis: The aim of the discourse analysis was to characterise the respondents' views on issues such as privacy and on the design options illustrated in the scenario mock-ups.

The interviewer's turns were composed of questions and explanations referencing the scenarios, privacy issues arising from monitoring, and socio-technical aspects of the system. Since the interviewer's discourse followed a planned script this is not reported in detail. The respondents' contributions were classified as shown in Table I.

Fig 2 Wmatrix semantic cloud of the post-interviews

TABLE I
RESPONDENT CONTRIBUTIONS

Questions of the interviewer	These were to do with *monitoring privacy* issues, clarification of the *scenarios* being presented, or *other*
Reaction to scenario Qs	Classified as *positive*, *negative* or *neutral*
Reaction to issue Qs	Also classified as *positive*, *negative* or *neutral*
Justification	For the responses given
Reflection	Classified as *general*, *personal history* (dementia experience, kin, etc.) or *self*
Computer experience	Classified as *general* (novice/expert), *use-specific episodes*, *devices and applications*, kinds of *activity.*
Other conversation	

The more frequent respondent categories were self reflection (10.8%), then other conversation (10.7%), followed by neutral reaction to issue (7.9%), then positive reaction to issue, computer device and question-other, all at 6.5%. Frequencies of other categories ranged from 1-5%. The net valency (positive minus negative reactions, ignoring neutrals) of reaction to the scenarios was +34 with the reaction to monitoring privacy issues slightly less favourable at +28. However, the group-level response masked considerable individual differences, as illustrated in Table II.

The two individuals who asked the most frequent questions (3 & 9) were also the most negative overall; however, the frequency of questions did not correlate with response valency overall. The response valency of issues and scenarios did correlate ($p<.05$, Spearman test), so people were consistent in their judgement. Respondents 3, 4, 5, 9 and 13 appear to be potential non-adopters of SAMS, whereas the rest appear to be moderate to strong potential adopters. These individuals also accounted for most of the privacy value concerns (72%), so invasion of privacy appears to be the main barrier to system acceptance.

The post-interview audio recordings were analysed with the categories listed in Table I, with the addition of further Reflection categories, on: *dementia*, *medical research on*

dementia, *healthcare/National Health Service (NHS) issues*, *news stories*, and *personal experiences*. The more frequent categories were personal history (17.9%), personal experiences (14.8%), reflections on news stories (9.5%), dementia (8.6%) medical research and healthcare issues both 6.3%. There were no obvious patterns in the data and few valenced reactions to privacy issues or SAMS.

TABLE II
DISTRIBUTION OF RESPONDENT QUESTIONS AND REACTIONS

Respondent no.	Questions	Net reaction-privacy	Net reaction-scenarios	Overall net valency
1	19	1	7	8
2	14	-1	7	7
3	43	0	-1	-1
4	10	-1	-2	-3
5	13	1	-4	-4
6	14	6	5	11
7	29	6	6	12
8	9	6	9	15
9	37	-1	0	-1
10	9	3	4	7
11	16	8	2	10
12	22	1	5	6
13	12	-1	-2	-3
Totals		+28	+34	

b) Frequency Profiling: The purpose of frequency profiling is to identify the most significant (not necessarily the most numerous) terms or concepts within a document. Here, we performed frequency profiling of semantic categories, to try to identify the dominant themes in the interview and post-interview transcripts. Note that in the manual discourse analysis, a tailored set of categories was defined (Table I) that was intended to classify the themes at the paragraph level. Wmatrix applies the USAS tags at the word or multi-word term level. Further, the USAS tags are general-purpose and do not map directly on to the tags in Table I. On the one hand, this makes it hard to use Wmatrix to directly validate the discourse analysis. On the other hand, it allows us to 'slice and dice' the transcripts in different but complementary ways.

Wmatrix was first applied to two documents: the consolidated interview transcripts and the consolidated post-interview recordings. As with the manual analysis, clear differences distinguished the two data sets.

These marked differences are illustrated by comparing Figures 1 and 2. Focusing initially on Figure 1, terms related to the solution domain (the tags *information technology and computing* and *Telecommunications*) occur frequently in the text. This suggests that more time in the interviews was spent exploring the design alternatives of SAMS, rather than respondents' emotional responses (e.g. the tag *worry*) or the problem domain (e.g. the tag *medicines and medical treatment*).

Terms tagged *information technology and computing (*e.g. *computer, laptop, website)* were the most over-represented category, with a log-likelihood (LL) of 704.74 representing a highly significant degree of over-representation relative to the British National Corpus's spoken English subset. This compared to an LL of 82.76 for terms tagged *medicines and medical treatment (*e.g. *doctor, diagnosed, blood-test)*, which still occurred significantly more often than predicted by the corpus, but less so than the IT and computing terms.

Note that there is always noise in such an analysis. For example, *pronouns* over-occur significantly, largely because the respondents frequently used first person personal pronouns when prompted to relate (e.g.) their preferences and experiences. It is because of this noise that Wmatrix is designed for supervised use. Clicking on one of the categories gives the concordances of all occurrences of terms tagged with that category, allowing the context of use to be verified. For example, clicking on *information technology and computing* gives direct access to all 215 instances of such words in the interview transcripts, with context, such as:

*I like the idea of being told to take an **online** test. So I quite like that. Just flash up.*

To filter out the noise, the individual respondents' contributions to the interviews were profiled (Figure 3) using the following subset of the most dominant semantic categories, represented by the USAS tags:

Y2	IT and computing
Q1.3	Telecommunications
A1.5.1	Using
X2.4	Investigate, examine
B3	Medicines and medical treatment
B2-	Disease

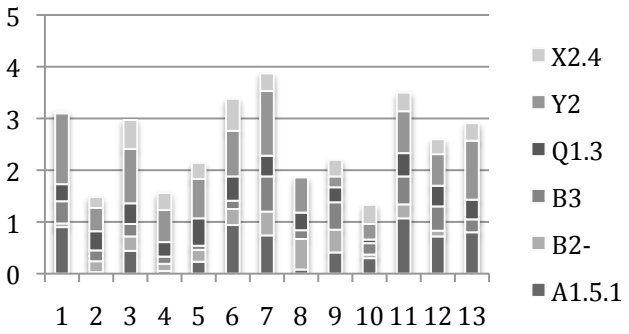

Fig 3 Respondents' interview contribution profiles

The y-axis represents the relative frequency of occurrence of terms tagged with one of the six tags in the respondents' interview transcripts. The overall mean for the cumulative relative frequency of the six tags was 2.53. The figure shows that respondents 2, 4, 8 and 10 contributed significantly less relative to the others in discussions on these themes. Respondent 10, for example, mentions e-mail only once, a term that is tagged with *Q1.3 Telecommunications*. This implied indifference to IT was interesting in itself, as one of the few occurrences of tag Y2 revealed in the concordance:

*I'm a bit anxious about trying any new gadget so I suppose my old **computer** would probably be the answer, as long as it doesn't interfere with any of the other programs.*

This was a useful contribution to one of the questions posed in the interview, which was whether, for the scheduled evaluation of the SAMS software, the respondents would accept a new laptop with the SAMS software pre-installed. We hoped for a 'yes' but it turned out that, like respondent 10,

many respondents found this unacceptable; they wanted to continue to use their own computers.

We also explored the privacy/security concern using the same technique. In this case, a different six USAS semantic tags were used, those that best corresponded to privacy and security:

A10-	Closed/Hiding
A15-	Danger
E6-	Worry/Concern
G2.1-	Crime
G2.2-	General ethics
S7.4+	Permission

As above, the tags were used to profile the respondents (Figure 4). The y-axis again represents the relative frequency of occurrence of the privacy- and security-related tags in the respondents' interview transcripts. The overall mean for the cumulative relative frequency of the six tags was 0.51. Thus, these tags were relatively over-represented in the responses of respondents 3, 5, 9, 10, 11 and 13.

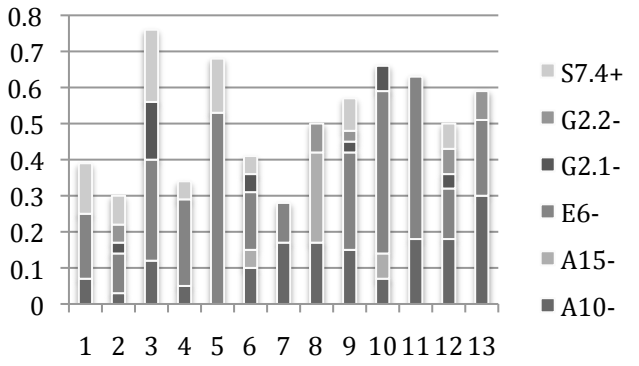

Fig 4 Respondents' privacy/security concern profiles

Each tag yielded a different set of terms in the transcripts. For example, E6 included *anxious*, *trust* and *reassuring*, while G2.2 yielded *misuse*, *sinister* and *immoral*. The relevance of the terms was easily verified by inspecting the concordances. For example, the following passage confirms the relevance of two instances of terms yielded by E6-:

*[illegal access] might have consequences as far as a financial sense I'm **concerned** or something like that, which I'm inclined to be **worried** about.*

The same tag (E6-) dominates the responses of 5 and 11 but their concordances reveal only general worry about dementia rather than about SAMS' implications for security and privacy. Respondents 3, 9, 10 and 13 thus appeared relatively concerned about privacy and security and this was broadly consistent with the findings of the manual analysis discussed earlier.

For the post-interview transcripts, problem domain issues (*medicines and medical treatment* and *disease*) predominate over solution domain (*information technology and computing*) as shown in Figure 2. The category *the media: newspapers, etc.* also occurs frequently, reflecting news stories providing cues for contributions.

2) Hypothesis-driven Analysis

The hypothesis-driven analysis sought to drill into the transcripts to find information related to what we believed would determine SAMS acceptance or rejection: the respondents' feelings and attitudes towards the system, as represented by their Values, Motivations and Emotions (VME) [33]. As before, manual and supervised semi-automatic analysis were performed, but this time the manual analysis was performed using eMargin [17] and the semi-automatic analysis was performed primarily to benchmark its performance.

eMargin is a collaborative annotation tool that allows the analyst to colour-mark passages of text, attach tags and share the marked-up document with other users.

The manual VME analysis investigated respondents' feelings and attitudes towards the system as well as providing evidence from emotions about possible acceptance or rejection of the system and the likelihood that system advice would be followed.

In the interviews the most commonly expressed emotion was anxiety (46.6 %), followed by distress and frustration (both 20%). Altruism, to take part in the research, was the most common motivation (52%), while privacy/security was the most frequently expressed value (72%). In the post-interview audio recordings, distress (35.2% of all emotions) and sadness (29.4%) emerged as the most frequent categories. However, motivations and values were rarely mentioned (4 and 7 total utterances), probably because the post-interview instructions biased the respondents towards reflection on their own experience and news stories.

Taking a high frequency of emotional expression to signify concern and hence an increased likelihood to adopt SAMS, all respondents apart from 4, 6, 7, 12 and 13 may be positively motivated towards the system. Respondent 13 had a poor reaction to the scenarios and low emotional engagement. Infrequent expression of emotion could indicate poor commitment to SAMS since these individuals are not motivated; whereas people who frequently express distress and sadness are probably better motivated to check their own health.

TABLE III
RESPONDENTS' FREQUENCY OF EMOTIONAL UTTERANCES

Respondent no.	Interview emotion	Post interview emotion	Net affect
1	0	6	6
2	5	1	6
3	5	3	8
4	0	No data	0
5	4	8	12
6	0	0	0
7	0	0	0
8	3	4	7
9	4	4	8
10	4	No data	4
11	3	7	10
12	1	1	2
13	1	No data	1
Totals	30	34	64

In the manual analysis, eMargin was used to annotate the text where respondents expressed values, motivations or emotions. The use of eMargin made the manual analysis available as a reference – a gold standard – against which the results of the Wmatrix analysis could be compared.

A subset of the USAS semantic tags were isolated that corresponded to VME. None were found to be satisfactory analogues for value or motivation, but the following were identified for emotion:

E3	Calm/Violent/Angry
E4	Happy/Sad
E5	Fear/Bravery/Shock
E6	Worry/Concern

In total, there were 275 instances of terms tagged E3, E4, E5 or E6, of which 75 corresponded to text marked up by the manual analysis, representing 27% precision. Only 25 of the manually marked-up passages were missed, giving recall of 75%.

C. Summary of Requirements

Reviewing the contributions of the manual and automated analysis, it is apparent that the 'traditional' manual analysis in section A identified the major user goals and non-functional requirements. However, only a few values, motivations and emotions were found even though the analyst was expert in such analysis and actively sought these insights. The more systematic manual discourse function analysis produced new findings on individual differences in user reactions to both the key NFR privacy and to design options in scenarios. A sub-group of users emerged (3, 4, 5, 9 & 13) who we considered likely to be unwilling adopters of SAMS. The manual VME analysis discovered another sub-group of users who showed little emotional reaction, which we considered unusual given the very real prospect of dementia affecting the lives of our senior citizen interviewees. We therefore interpreted users (4, 6, 7, 12 & 13) as another group of unwilling adopters.

The text mining, described in section B found another user sub-group of low responders (2, 4, 8 & 10) who, apart from respondent 4, did not intersect with the other two groups of potential unwilling adopters. We conjecture that low frequencies of dominant semantic categories, when these are oriented towards the solution domain, might indicate a more benevolent attitude to the system. These individuals also asked fewer questions in the interviews. Text mining of semantic categories corresponding to the privacy/security concern, supported the manual analysis identification of the privacy-sensitive user sub-group. Manual analysis of emotion was also confirmed by text mining the corresponding categories with a recall of 75%. These results give us confidence that automated text-mining tools can replicate a manual analysis even in difficult areas of sentiment. The automated analysis also produced tags that posed further analysis questions when tags were followed to utterances in the source text, as illustrated in the Computing Technology concerns.

The systematic manual analysis, excluding transcription time, took approximately 45 hours for coding. The informal audio-only analysis took 18 hours listening to interview recordings while making notes. In contrast, automated analysis was rapid, accounting for 6 hours including data cleaning and producing the results. The text-mining tools therefore afforded a considerable time saving while producing results of similar quality. As data volumes scale up this advantage will become more significant.

The requirements implications of the user sub-groups lies in customisation. For the privacy-resistant sub-group we will restrict monitoring to quantitative measures in e-mail text, so no semantic interpretation is undertaken. This, in combination with a user control to temporarily disable monitoring, will be explained to reassure this sub-group. The low emotion sub-group will need increased motivation. Here we will employ video explanation and persuasive technology design guidelines [11], e.g. use of praise, personal address, etc. to increase user commitment to SAMS.

VI. Discussion

A. Threats to Validity

Internal threats to the validity of the analysis were minimised by having the first author running the workshops and interviews and the subsequent manual analysis, while the other three authors variously applied the Wmatrix tool. This meant that the transcripts were processed blind by the tool users, avoiding bias regarding what to look for and where.

Some of the themes that emerged during the workshops, interviews and post-interview sessions conflated requirements issues for research and for what we hope will eventually be a deployed system. For example, altruism was a motivation of the respondents who volunteered to provide requirements information for SAMS, but their altruism was directed towards the research project and was not indicative of whether an altruistic person would be more or less likely to use the SAMS software. This confusion was hard to eliminate and adds noise to the already difficult problem of interpreting the data on affect.

B. SAMS Requirements

The apparent correlation between questioning and negative valence may offer an insight into who may and may not be potential users of SAMS. Similarly, we have hypothesised that the people who articulated more emotion are more likely to adopt SAMS. The affect analysis elicited user sub-groups that have implications for customisation; reassure one group worried about privacy and increase the motivation of the other low emotion group. Text mining for user profiles may have further potential in personal RE [31], for instance in eliciting individual user goals. Clearly there were limitations on the applicability of text mining. For example, apart from privacy, automated identification of motivations and values was not possible since few explicit lexical markers of these categories appear in text. Despite these caveats, the exercise did reveal important requirements about privacy and alert types that may be characteristic of senior users.

C. Methodology

The phased elicitation sessions served the important purpose of allowing us to focus on issues that emerged in earlier sessions. The interviews helped explore issues that emerged in the workshops. The self-recorded sessions, were intended to encourage respondents to provide thoughtful

insights outside the structure of a formal elicitation event, and they were somewhat successful in providing information about affect.

The analysis to which we subjected the recorded text provided some interesting data with one unanticipated correlation between questioning and negative valence. This was time-consuming work, even for the relatively modest number of respondents, and it is possible that further tool support, such as sentiment analysis to generate the valence results, may be of use here in future.

The observed performance of Wmatrix needs to be treated carefully. At first glance, frequency profiling appeared to give very different results from the manual discourse analysis. This was because of the differences between the tailored set of categories used in the discourse analysis and the general-purpose nature of the USAS tags. The fact that Wmatrix is applied to individual terms or multi-word units, while the discourse analysis is classified at the paragraph level, also contributes to the difference. However, by providing an alternative way of viewing the text, purely as a data set, frequency profiling was able to complement the manual analysis, finding different ways to contrast the respondents' contributions, allowing the analyst to follow the evidence or even serendipitously following their own hunches into the text by clicking on tag terms to view a set of concordances.

Filtering on tag subsets that map on to particular classes of information is limited by the general-purpose nature of the tagset. No tags mapped on to value or motivation, for example. Where analogues were found, however, performance was reasonable and suggests that there is potential for similar tools.

With respect specifically to identification of affect-laden utterances, if the observed level of performance (27% precision, 75% recall) proved to be replicated in future analyses of affect, it would be tolerable provided the following caveats held:

- The assessment was done across a sufficiently wide and representative user population; and
- Developing an understanding of the *range* of emotional responses was more important than collection of *all* responses.

Thus, rapidity of analysis would compensate for having only qualified confidence in the result. Such a trade-off is inherent for applications of text mining but it means that text mining is not suitable in all RE contexts. In particular, it should be noted that while easy access to terms' context in the text using concordances makes validation of true positives and rejection of false positives easy, only a painstaking manual analysis can reveal the false negatives. If this needs to be done, it negates any effort-saving advantages of text mining [4]. Nevertheless, text mining did allow us to posit heuristics which may direct future automated analysis of affect-laden applications, e.g. screen users for negative emotional reactions, low frequency of emotions, and sensitivity to key values.

For comparison, with the privacy and security tags (Figure 4), 63 of the 172 instances of terms returned by the tags proved to be of relevance, giving a precision figure of 37% when we checked each instance. Since eMargin mark-up had not been used for the data-driven analysis, we were unable to quantify recall.

The implication of this is that, of the responses plotted for each respondent in Figure 4, only a third are likely to be a true positive, i.e. relevant to understanding the privacy/security concern. However, since each column is made up of the same tags, and if we assume that the precision was consistent across each tag for each respondent, the conclusions we drew from the data in the graph remain robust.

As usual for text mining techniques, there is an inverse relationship between recall and precision; recall can be increased by including more tags, but at the expense of precision. In RE applications, recall is generally favoured over precision, since errors of omission are harder to detect than errors of commission. Seen in this light, the performance of Wmatrix was adequate, permitting hypotheses to be checked quickly.

A final reflection concerns the scenario used for text mining directed RE. As applications come to be delivered more frequently over the web, and as requirements are also analysed by remote web capture, the collection of user feedback to mock-ups and prototypes may become commonplace. Given high numbers of potential users and corresponding diversity of needs, automated analysis of linguistic feedback, opinions and affect will begin to pay off.

VII. Conclusions

We set out to develop a tailored requirements discovery method that addressed critical questions of technology acceptance for affect-laden problems. We also wanted to exploit tool support to help our analysis. In the end we gained insights into the myriad ways in which affect colours peoples' views of health-related problems and systems, and we gained insights into how seniors regard technology as risky, and their concerns for privacy and security. While the implications for customisation we have drawn from this analysis will need to be validated by user reaction to SAMS in practice, the process we have undertaken has potential for application in other domains where motivation and affect are important.

We have accumulated information that has allowed us to better understand the requirements for SAMS. We have also gained insights into the potential and limitations of shallow semantic analysis for RE.

In the next phase of the project, we will be developing the SAMS software and subjecting it to trials that are intended to reveal more about the extent to which it will prove acceptable before deploying SAMS in a longitudinal study of real users.

Acknowledgments

The work described in this paper is funded by EPSRC project ref. EP/K015796/1 *Software Architecture for Mental Health Self Management* (*SAMS*).

References

[1] Alwan, M., Mack, D.C. et al. (2006), "Impact of passive in-home health status monitoring technology in home health: Outcome pilot" in *Proceedings 1st Transdisciplinary Conference on Distributed Diagnosis and Home Healthcare, D2H2-06*. pp.79-82.

[2] Ballard C. and Corbett A. (2012), "Screening for dementia: An opportunity for debate", *Expert Rev Neurother*. vol. 11(10), pp. 1347-9.

[3] Beckles, B. (2005), "User requirements for UK e-Science grid environments" in *Proceedings 4th UK All Hands E-Science Meeting*.

[4] Berry, D., Gacitua, R., Sawyer, P., Tjong, S.F. (2012), "The case for dumb requirements tools" in *Proceedings 18th International Working Conference on Requirements Engineering: Foundations of Software Quality (REFSQ'12)*, pp 211-217.

[5] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent dirichlet allocation", *Journal of Machine Learning Research*, vol. 3, pp 993-1022.

[6] Callele, D., Neufeld E. and Schneider, K. (2009), "Augmenting emotional requirements with emotion markers and emotion prototypes", in *Proceedings, RE 2009*. Los Alamitos CA: IEEE Computer Society Press, pp. 373-374.

[7] Cleland-Huang, J., Berenbach, B., Clark, S., Settimi, R., Romanova, E. (2007), "Best practices for automated traceability", *IEEE Computer*, vol. 40, pp 27-35.

[8] Cysneiros, L.M., (2002), "Requirements engineering in the health care domain", *Proceedings 10th IEEE Int. Conf. on Requirements Engineering*. Los Alamitos CA: IEEE Computer Society, pp. 350-356.

[9] Davis, F.D. (1989), "Perceived usefulness, perceived ease of use, and user acceptance of information technology", *MIS Quarterly*, vol. 13(3), pp. 319-340.

[10] Dumais, S., Furnas, G., Landauer, T., Deerwester, S. and Deerwester, S. (1995), "Latent semantic indexing", *Proceedings. 4th Text REtrieval Conference (TREC-4)*.

[11] Fogg, B.J. (2003), *Persuasive Technology: Using Computers to Change What We Think and Do*, San Francisco: Morgan Kaufmann.

[12] Gacitua, R., Sawyer, P. and Gervasi, V. (2011), "Relevance-based abstraction identification: Technique and evaluation", *Requirements Engineering*, vol. 16(3), pp. 251-265.

[13] Garde, S. and Knaup, P. (2006), "Requirements engineering in health care: The example of chemotherapy planning in paediatric oncology", *Requirements Engineering*, vol. 11, pp 265-278.

[14] Hayes, J.H., Dekhtyar, A., Sundaram, S.K. (2006), "Advancing candidate link generation for requirements tracing: The study of methods", *IEEE Transactions on Software Engineering*, vol.32, pp.4-19.

[15] Jirotka, M., Procter, R. et al. (2005), "Collaboration and trust in healthcare innovation: The eDiaMoND case study", *Journal of Computer-Supported Cooperative Work*, vol. 14(4), pp. 369-398.

[16] Jorgensen, J.B. and Bossen, C. (2003), "Requirements Engineering for a pervasive health care system" in *Proceedings 11th IEEE International Conference on Requirements Engineering (RE'03)*. pp. 56 – 64.

[17] Kehoe, A. and Gee, M. (2012), "eMargin: A collaborative text annotation tool." in *Proceedings 33rd Conference on International Computer Archive of Modern and Medieval English (ICAME 33)*.

[18] Le, X., Lancashire, I., Hirst, G. and Jokel, R. (2011), "Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists", *Literary and Linguistic Computing*, vol. 26(4), pp. 435-461.

[19] McArthur, T. (1981), *Longman Lexicon of Contemporary English*. London: Longman.

[20] Rodríguez-Ferreiro, J., Davies, R, González-Nosti, M., Barbón, A. and Cuentos, F. (2008), "Name agreement, frequency and age of acquisition, but not grammatical class, affect object and action naming in Spanish speaking participants with Alzheimer's disease", *Journal of Neurolinguistics*, vol. 22(1), pp 37-54.

[21] Nasreddine, Z. et al. (2005), "The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment". *Journal of the American Geriatrics Society*, vol. 53(4), pp. 1532-5415.

[22] Perrone, V., Finkelstein, A. et al.. (2006), "Developing an integrative platform for cancer research: A requirements engineering perspective", *Proceedings. 5th UK All Hands e-Science Meeting*.

[23] Pilemalm, S. and Timpka, T. (2008), "Third generation participatory design in health informatics: Making user participation applicable to large-scale information system projects". *Journal of Biomedical Informatics*, vol. 41, pp. 327-339.

[24] Popescu, D., Rugaber, S., Medvidovic, N. and Berry, D.M. (2008), "Reducing ambiguities in requirements specifications via automatically created object-oriented models" in Paech, B., Martell, C. (eds) *Innovations for requirement analysis: From stakeholders' needs to formal designs*. Heidelberg: Springer, pp. 103–124.

[25] Ramos, I. and Berry, D.M. (2005), "Is emotion relevant to requirements engineering?", *Requirements Engineering*, vol. 10(3), pp. 238-242.

[26] Ramos, I., Berry, D.M. and Carvalho, J. (2002). "The role of emotion, values and beliefs in the construction of innovative work realities" in *Proceedings of SoftWare 2002: Computing in an Imperfect World*.

[27] Rayson, P. (2008), "From key words to key semantic domains", *International Journal of Corpus Linguistics*, vol. 13(4), pp. 519-549.

[28] Rayson, P., Archer, D., Piao, S.L., McEnery, T. (2004), "The UCREL semantic analysis system", In *Proceedings of the Workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks, in association with 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 7-12.

[29] Ryan, K. (1993), "The role of natural language in requirements engineering" in Proceedings 1st Int. Symposium on Requirements Engineering. Los Alamitos CA: IEEE Computer Society, pp. 240-242.

[30] Sawyer, P., Rayson, P. and Cosh, K. (2005), "Shallow knowledge as an aid to deep understanding in early-phase requirements engineering", *IEEE Transactions on Software Engineering*, vol. 31(11), pp. 969-981.

[31] Sutcliffe, A.G., Fickas, S. and Sohlberg, M.M. (2006). "PC-RE: A method for personal and contextual requirements engineering with some experience," *Requirements Engineering*, vol. 11, 157-163.

[32] Sutcliffe, A.G. and Sawyer, P. (2013), "Requirements elicitation: Towards the unknown unknowns" in *Proceedings 21st IEEE International Conference on Requirements Engineering*, pp. 92-104.

[33] Sutcliffe, A.G., Thew, S. et al. (2010), "User engagement by user-centred design in e-health", *Philosophical Transactions of the Royal Society, special issue on e-Science, series A*, vol. 368, pp. 4209-4224.

[34] Thew, S. and Sutcliffe, A.G. (2008). "Investigating the role of soft issues in the RE process" in *Proceedings 16th IEEE International Requirements Engineering Conference*, pp. 63-66.

[35] UK Department of Health (2013), "G8 dementia summit communique" https://www.gov.uk/government/publications/g8-dementia-summit-agreements. Accessed 10 March 2014

[36] Wilson, T., Wiebe, J. and Hoffmann, P. (2005), "Recognizing contextual polarity in phrase-level sentiment analysis" in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 347-354.