The language of information need: differentiating conscious and formalized information needs

Ian Ruthven

Department of Computer and Information Sciences, University of Strathclyde, Glasgow, United Kingdom

Abstract

Information need is a fundamental concept within Information Science. Robert Taylor's seminal contribution in 1968 was to propose a division of information needs into four levels: the visceral, conscious, formalized and compromised levels of information need. Taylor's contribution has provided much inspiration to Information Science research but this has largely remained at the discursive and conceptual level. In this paper, we present a novel empirical investigation of Taylor's information need classification. We analyse the linguistic differences between conscious and formalized needs using several hundred postings to four major Internet discussion groups. We show that descriptions of conscious needs are more emotional in tone, involve more sensory perception and contain different temporal dimensions than descriptions of formalized needs. We show that it is possible to differentiate levels of information need based on linguistic patterns and that the language used to express information needs can reflect an individual's understanding of their information problem. This has implications for the theory of information needs and practical implications for supporting moderators of online news groups in responding to information needs and for developing automated support for classifying information needs.

Keywords: information needs, internet forums, emotion, cognition, sensation

2010 MSC: 00-01,99-00

1. Introduction

The concept of information need has been fundamental to many models and studies within Information Science and has been the motivation for much research on information seeking, information use and interactive systems design. In his seminal work on information needs, Taylor proposed that information needs exist across four levels: the visceral, conscious, formalized, and compromised levels [1]. The visceral need is the unexpressed need that may only reflect a '*vague feeling of dissatisfaction'* one that is *'probably inexpressible in linguistic terms'* and, as Cole later put it, '*unspecifiable even to the user herself*'[2] The conscious need is '*a conscious mental description of an ill-defined area of indecision*' that results from the conscious recognition of a problem that requires attention. The formalized need is '*a qualified and* *rational statement of his question*', i.e. a clear expression of an information need. Finally, the compromised need is the question that is posed to the (human or technical) information system.

Although Taylor seemed concerned that what he was stating what was obvious, at least to library professionals, his conceptualisation of information need levels has been immensely influential and his paper has become one of the most cited works in Information Science [3]. It has influenced research within Information Science, including work on system design, e.g. [4], and information behaviour studies, e.g. [5], and outside of Information Science, e.g. the work of Jansen who used Taylor's four levels to model the process of selling online [6].

Chang's recent citation analysis of Taylor's paper demonstrated that the four levels of information need has been the most cited contribution from this work [7]. This includes a recent and very considered article by Cole, who proposed a re-interpretation of Taylor's four levels, not as stages in information need development, but as different levels of understanding of the same need [8]. The idea that different levels of need have different characteristics has also been supported by Lundin [9] who sees the compromised need level as one way of framing desired goals whilst Agarwal [10] sees the visceral level being the area where serendipity is most likely to occur.

In spite of decades of work on information need, it is still a concept that is poorly understood [11-13] with Savolainen noting in a recent review '*even though information need is probably the most widely used construct explaining why people engage in information seeking, this concept is still vague*' [12]. Lundh posits that one reason for this lack of specificity around what we mean by information need is that information needs, as internal cognitive states, are difficult to investigate directly and it is easier to investigate the information behaviours and activities arising from information needs instead [14].

Taylor's division of information needs is described typically in conceptual or anecdotal terms with the division taken as a useful axiom about information needs but rarely analysed itself. In preparing for the work in this paper, we analysed the 300 odd hundred papers that cited Taylor's work in the five preceding years, (January 2013 to April 2018) and none provided, or cited, any empirical way to differentiate between the levels. That is, we lack empirical investigations on the difference between levels of need that go beyond Taylor's original conceptual work and that can be used in practical settings. In this paper, we conduct a linguistic analysis of written information need statements that demonstrates we can differentiate between different levels of information needs based on the language used to describe the needs.

Our proposal is that people reveal their inner states when posting needs to online forums and by analysing these posts we can identify what level of information need they are experiencing. As well as providing a new understanding of how information needs differ, and therefore how the response to such needs may differ, this also opens up the opportunity to detect automatically what level of need is expressed in requests for help made online and therefore how moderators of discussion groups should respond to posts.

We consider first the related literature to motivate four hypotheses on differences between information need statements that correspond to Taylor's conscious and formalized information needs. Following this, we present a series of empirical investigations followed by a discussion of our findings and their implications for future research.

2. Literature Review

In this section, we consider four areas of relevant literature on information needs. The literature on information needs is vast and so here we restrict our analyses to contributions that were useful to develop our hypotheses. We particularly focus on the differences between Taylor's conscious and formalized information needs as the two need types that are most amenable to linguistic analyses (see section 3.2 for more on this point).

2.1 Information needs and problematic situations

Even though information needs are a core concept in information seeking, it is not a given that information needs are what we should be investigating when we study information seeking. Other authors have proposed that the situations that require information are a better focus of study. Taylor himself noted that '*inquirers frequently cannot define what they want, but they can discuss why they need it'*, i.e. people often cannot say what is their information need but can talk about the situations that have given rise to information needs [1].

Others have also commented on the issue that people do not think in terms of formal statements of need but rather on problematic situations that have to be turned into expressions of information need in order to obtain information. For example, Belkin et al. in their famous contribution on Anomalous States of Knowledge (ASK) stated '*information need is in fact not a need in itself, but rather a means toward satisfying some more basic need, typically, in the situations with which information science is concerned, the resolution of a problem*' [15]. They also proposed that this situation-based understanding of need should affect how we design information systems '*for our representation of ASKs are narrative statements by the users of the IR system, of the problems which brought them to the system*' and therefore are better ways for users to present needs to systems than queries.

These authors placed an emphasis on the *situation* where information might be useful rather than on the information *need* itself. This situational view is still a popular approach to understanding information behaviour, see for example the discussions in [13, 16, 17].

People in problematic situations have the challenge of either turning their situation into a need statement (translating their conscious need into a formalized one) or presenting their situation to someone who can help this transformation. Problematic situations are closest to Taylor's conscious level of need, which he felt would need dialogue with someone else to clarify. In his '*ambiguous and rambling*' depiction of conscious information needs, Taylor noted that conscious needs might lack focus due to their emerging nature. Similarly, Belkin et al.'s description of '*narrative statements…of the problems that brought them to the system'* suggests that early stage information needs are more likely to be descriptive accounts of a problem situation rather than a precise expression of a need. Formalized needs though, from Taylor, are '*qualified and rational statement of his question*' – the use of the word 'question' suggesting that the situation has resolved into a need statement.

This division between narrative descriptions of problematic situations (conscious need level) and focused descriptions of need statements (formalized need level), leads us to hypothesize that textual descriptions of conscious information needs will be longer than those of formalized needs due both to the need to describe a situation and the lack of an ability to express a precise (formalized) need.

Hypothesis 1: We hypothesize that statements describing conscious information needs will be longer than statements describing formalized information needs.

2.2 Information needs and emotion

Information needs and uncertainty are tightly linked [18]. Even though uncertainty can have positive dimensions, such as excitement or curiosity [19], uncertainty within information seeking research has typically been connoted with negative emotions [18, 20, 21].

Kuhlthau [22] in her seminal work on the Information Search Process (ISP), and directly influenced by Taylor, tackled the issues of uncertainty and emotion, noting that information can be disruptive and cause confusion and distress rather than comfort and certainty. Uncertainty can cause negative emotions in early stages of information need development, as Kuhlthau [22] states '*uncertainty, a natural and necessary aspect of the early stages of the ISP, causes discomfort and anxiety which in turn affects articulation of a problem*', also observing that '*an inability to express precisely what information is needed*' co-occurs with '*uncertainty, confusion and doubt'*. In later stages of the ISP, when needs become focussed '*a change in feelings is commonly noted, with indications of increased confidence and sense of clarity'* [22]. Therefore, the emotions relating to early stage information needs are more negative; the ones relating to later stage needs, more positive.

Similarly, Braschers et al. [23] refer to '*ambiguous situations ... cause many anxious times'* and Taylor talks about the earliest stages of information need (visceral need) as being a '*vague feeling of dissatisfaction'*. Nahl and Bilal [24] talk about affective uncertainty *as 'a feeling of unease due to the presence of cognitive uncertainty and it can be experienced as irritation, frustration, and anxiety'* and Zhang [25] describes the various '*emotional motivations*' for engaging in online information seeking as including '*uneasy and disturbing feelings about conditions of themselves or of people who they care about'* noting that '*some [people] felt embarrassed, troublesome, nervous, worried, upset, and anxious; while others felt miserable, desperate, going crazy, freaking out, and scared to death*.'.

Therefore, we see a strong link between greater uncertainty and greater emphasis on negative emotions, particularly anxiety. In Taylor's conscious need level there is greater uncertainty about the situation being faced and what may help. We might, therefore, expect that even if people cannot express what information they want, they can express what they *feel* about a situation and that negative emotions will be more commonly expressed to describe conscious needs, where we do not know what information we need, than formalized ones, where we can be confident about our information need.

Hypothesis 2: We hypothesize that statements describing conscious information needs will contain an increased use of negative emotional words compared to statements describing formalized information needs.

2.3 Information needs and sensation

Increasing attention is being paid in the Information Science literature to the body. This research comes from two directions. The first focusses on embodiment, investigating the body as a source of information for cognition, e.g. [26-28]. As Olsson and Lloyd [27] explain, there is a substantial literature from outside of Information Science that recognises the importance of the body for '*the demonstration of practical reasoning*' but Information Science has, until now, seen far less development in this area.

Sensory activities can provide important information for professionals with Olsson and Lloyd writing about how nurses' '*Sensory activities such as touching and smelling represent critical activities'*. Sensory properties can also be useful for improving lay people's interpretive abilities about their own bodies. Brashers et al. [29] for example noted how people with chronic illness, over time, can develop strong interpretive abilities about their own bodies and the significance of physical signs which may or may not indicate a health concern.

Similarly, Godbold [30], looking at the interactions and discussions of people of people on a renal support bulletin board, noted how often the body and interpretations from bodily experience can be an important source of information, '*or that they knew there was a problem because of a sensation they noticed'*. Godbold also observed how participants on the bulletin board '*used measurements and sensations as informative elements that they brought together to justify or question how they understood situations*.' One of Yates' [28] seven frames of health literacy is '*paying attention to bodily information'* where literacy means watching for changes to physical states. Information itself is described as physical changes including '*unpleasant physical sensations (e.g., pain, discomfort), changes in people's physique, or bodily reactions that are perceived as different or unusual.*' This awareness of change is the first stage in trying to uncover what the change means.

A second focus for the body is the idea of intuition and the physical sensations that we use to determine when a situation needs closer attention. In many areas of life, intuition based on sensation is an important way of interacting with the world. In their review of the literature, Douw et al. [31] looked at various sources of intuition within nursing practice; signs that something was not right and that a nurse should be concerned. These included nurses own qualitative evaluations such as *'[patient] does not look or seem right', 'something is not right'* and *'a look in the eyes [that indicates something is different]*'.

King and Clark [32] also investigate the power of intuition stating that '*intuitive awareness appeared to become an increasingly powerful aspect in some of these nurses' decision-making. It appeared to act as a trigger, sparking an analytical process that involved the nurses in a conscious search to acquire data that would confirm their sense of change in the patient's status*', arguing that our physical gut-feeling can be the first step in understanding whether something is wrong and leading to seeking information to better understand a situation and what is required in that situation. Taylor himself called his first level of need the visceral level signifying a physical awareness of need as being important.

If we are struggling to understand our situation, and possibly whether we are in a problematic situation at all, then we may have to rely more on sensory signals as a source of understanding our situation. That is, we may be in a position where we have to work with what we can describe (our sensations and feeling) than what we cannot describe (the information that may be necessary to resolve our situation). Therefore, we expect to find more use of words relating to perception and sensation in early information needs.

Hypothesis 3: We hypothesize that statements describing conscious information needs will contain an increased use of words describing physical sensations compared to statements describing formalized information needs.

2.4 Information needs and cognition

As described in section 2.2, conscious information needs are ones that involve more uncertainty. Brashers [33] propose that '*Uncertainty exists when details of situations are ambiguous, complex, unpredictable, or probabilistic; when information is unavailable or inconsistent; and when people feel insecure in their own state of knowledge or the state of knowledge in general".*

In problematic situations, uncertainty is not necessarily restricted to one aspect of a situation. Brabow [34] points out that we may have multiple uncertainties at the same time, they may interact and have different valences, resulting in us "*chaining from one focal dilemma to another*". Brabow points to communication as a way of resolving uncertainties (but also potentially making them worse). Taylor also notes that communication in the conscious stage can help in that someone else may understand the '*ambiguities'* in the situation and that these ambiguities '*will gradually disappear in the course of the dialogue'*. '*Disappearing'* ambiguities allow the person to form '*qualified and rational statements'* in the formalized need level [1]. The compromised level of information need suggests a level of need where one knows exactly what to expect from the system; the lower levels have an uncertainty about the information that may be obtained and the form it may take.

Lundh proposed an interesting difference between various information need levels, describing visceral/conscious needs as information needs and formalized/compromised needs as 'questions', suggesting that the last two levels are somehow qualitatively different from the first two levels [14]. Taylor himself, and others such as Cole [8], support Lundh in reminding us that there is not a development from a conscious need into *a* formalized need but rather there are potentially many and different formalized needs depending on the situation being tackled and how we are able to understand and pose questions about it. Ingwersen, in discussing what he referred to as the 'labelling' effect of having to describe a formalised need, reports that '*This labelling effect often misrepresents the subject area needs to the intermediary, and thus the label may well fall outside the context of the user's real need*.' [35], expressing the concern that the labelling effect may lead to different understandings of the real need by intermediaries and possibly different interpretations by different intermediaries.

Sometimes a situation may be easily resolved into a single statement of need; other times we may need to ask questions to probe what information is available or to gain more information to better understand our situation. However, moving into the stage of creating formalized needs may help us understand what information we possess and begin the process of knowledge construction around our problem area. This is similar in spirit to Kuhlthau's exploration stage in her ISP characterised by doubt, uncertainty and confusion and where her participants had an '*inability to express precise information needed'* but were '*intentionally seeking possible focuses'* to move the problem forward [36]. The visceral information need that underlies our search process may only be revealed slowly, and in parts, and which areas of the visceral need are revealed may depend on *how* we move from visceral to conscious to formalised and compromised needs. Therefore, our process of knowledge construction may lead to different outcomes even when starting from the same visceral need.

The process of moving from simply being able to describe our problematic situation (conscious need) to being able to ask questions about it (formalized need) requires active cognition to think about the problem in sufficient detail to move the problem forward. Therefore, we might expect that in conscious need level there are more words expressing active thinking about a problem and what may help in that situation whereas in formalized information needs we have already moved to knowing what information we require and therefore our descriptions of these needs involve fewer 'thinking' words.

Hypothesis 4: We hypothesize that statements describing conscious information needs will contain an increased use of words describing cognition compared to statements describing formalized information needs.

3. Methods

3.1 Data

A common method for resolving information needs is to go online and interact with people in discussion groups or online forums. Online interactions can put us in touch with people who have experienced similar situations to us and who can emphasise and offer advice and information [37-39].

Online forums can also help us understand how Taylor's levels of information needs differ by providing textual descriptions of people's perceived needs. There are many advantages to using online data. Firstly, the needs are expressed as they are felt. That is, people post on what is of current concern, generally as narratives, rather than later reflections on experiences. This means the posts are 'immediate' in providing the context of a person thinking through their situation. Secondly, we can deal with many more people and their stories than is typically reported through interviews and other narrative forms. This allows for more experience to feed into the analysis and for more minority experiences to be involved. A major advantage is that the needs are described textually: they are written statements of a situation and/or need that can be analysed textually to uncover patterns within the texts.

Set against this there are disadvantages. Unlike surveys or interviews, we cannot ask questions of the participants; neither can we clarify meanings or follow up later. We can only work with what is expressed rather than the totality of what is felt of experienced. Nevertheless, as we show in this paper, even with these limitations, textual analyses can be very powerful in differentiating information need levels.

In this paper, we use data available from online forums to investigate the differences between Taylor's conscious and formalized information need levels. We focus on these two need levels as visceral needs, according to Taylor are '*probably inexpressible in linguistic terms'* and so unlikely to be asked about in online forums which require a linguistic description. Compromised information needs are needs expressed in order to gain information from a specific information system and expressed in terms of that information system. In the case of online forums, there is no correspondence to such information systems, beyond perhaps the choice of which forum to choose, and so this case does not occur in our data. The space between the conscious and formalized needs is, however, of particular interests as, according to authors such as Lundh, it marks the point where internal needs become external, linguistic entities [14].

We use four datasets in this paper, each created from a popular UK-based forum devoted to a distinct area of life. We selected datasets on different topics to uncover more generalizable patterns than may be possible when only looking at one domain. For each dataset, the use of the word 'post' refers to the first post in a discussion, the post that contained an information need. We analysed no responses to posts. In line with generally agreed ethical standards for conducting online research, we only examined posts from major forums that do not require registration to view, from groups which are more likely to be considered 'public spaces' and which have more than 100 members [40].

For all forums, we removed requests made on behalf of another person as we wished to analyse personal information needs. We also removed all opinion and speculative questions, e.g. *what do people think Bitcoin will be worth in a year's time*, if they seemed to be intended to start a discussion rather than answer a personal need. We retained questions that asked for opinion if they seemed they were asked in order to resolve some uncertainty, e.g. asking whether a symptom was unusual.

These forums are used to request information but also for other purposes, e.g. sharing news items or distributing surveys, status updates on people's lives, etc. These were all removed and the remainder of posts were checked for the presence of an information need. As noted above, with forum posts we cannot clarify with the original poster their intention of posting to a forum. Rather, we can only estimate the most likely reasons for their posting from the text and from any available responses to the post: are they seeking information, are they making an announcement of their situation or simply emotionally 'venting' [41].

To test this classification – as to whether a post contains a personal information need – we asked a colleague not associated with the work to classify a randomly selected 10% sample of the posts. The inter-coder reliability test showed a Cohen's Kappa coefficient of 0.75, substantial agreement strength on whether the post contained an information need [42].

Our first dataset is the **Diabetes** dataset, a sample of posts from the Diabetes UK "Diabetes Support Forums". Diabetes UK is a major UK charity providing support for people with diabetes and funding into diabetes research. We took an eight-week sample from this site, consisting of all posts from 7th Feb 2017 to 4th April 2017, providing 585 posts. Once we applied the exclusion rules described above, we had a set of 323 posts for analysis.

The second dataset is the **Finance** dataset, a sample of posts from The Student Room "Money and Finance" forum. The Student Room is a student community and forum website aimed at those who are considering going to University and those already at University in the UK. We took a twelve-week sample consisting of all posts from 16th August 2017 to 21st November 2017. Applying our exclusion rules left a dataset of 268 posts for analysis.

The third dataset is the **Mothers** dataset consisting of posts from young (less than 21 years old) first-time mothers. This consists of all posts from the NetMums' "Young Parents Support" forum and the BabyCentre forum, posted from mid-August 2014 to mid-August 2015, that met our inclusion rules, providing a dataset of 266 posts.

The fourth dataset is the **Sexuality** dataset, a sample of posts from The Student Room "Sexual Health" forum. We took a four-month sample consisting of all posts from 7th December 2016 to 3rd April 2017 which, after filtering through our exclusion rules, left a dataset of 292 posts.

3.2 Classification into conscious and formalized needs

Ruthven et al. [43] showed that posters to online groups often present information needs at different levels with a classification of posts into what they referred to as Informational and Situational posts. These were described as corresponding to Taylor's conscious and formalized information needs and distinguished between posts where the poster could identify what information she needed (Informational needs) and cases where the poster could not yet formalize her need into an information need (Situational needs).

Following this approach, we took each post in each dataset and classified it into either being a **conscious** or **formalized** statement of need. We used Taylor's conceptual description of conscious or formalized needs as a guide to inform our coding process. In his paper, these two need types are regions on a spectrum: conscious needs can be at a very early stage and close to Taylor's description of visceral needs where the poster is only aware of a '*vague sort of dissatisfaction'* [1] and is starting to question whether he is in a situation that needs information. Later conscious needs can be very close to formalized needs where the poster is already starting to make sense of her situation and is close to identifying what kinds of needs she has.

The key distinction for our classification was the presence of a clearly stated expression of information need. These were classified as posts containing **formalized** needs**.** These needs come from people who, in Taylor's words, can '*form a qualified and rational statement'* and who, if necessary, can describe '*his area of doubt in concrete terms*'. Examples include *'...he* [poster's baby] *is also waking up several times a night hungry. does this mean his system is ready for something more substantial?*', '*Can anyone tell me if diabetic Easter eggs are no good for you or is it better. to,have a dark one*.' and '*I am just applying for student finance and would like to know the best bank to sign up to*.'

The other posts were classified as **conscious** posts**.** Conscious needs may contain questions but these questions will be vague and often of the form '*Has anyone been in this situation or am I being silly'*, '*I've got to the point where I don't know what to do, if you were in my position what would you do'*, or *'does anyone have any advice for me?*' where people require help but it is unclear what form a suitable answer will take. These posts correspond to Taylor's description of conscious information needs as being '*ill-defined area[s] of indecision'.*

We conducted a test of this classification by asking a second colleague, not involved in creating the classification, to classify a randomly selected 10% of the posts. This sample consisted of every 10th post within our datasets to provide a representative sample of the data and post types. The inter-coder reliability test showed a Cohen's Kappa coefficient of 0.68, good agreement strength on which types of information need were contained within posts [42].

Table 1 summarizes the division of the datasets into conscious and formalized needs. The distribution of conscious to formalized needs varies between datasets but both types of needs are present in each dataset.

Table 1: Distribution of conscious and formalized needs across the four datasets

	conscious	**formalized**
diabetes	149 (46.1%)	174 (53.9%)
finance	88 (30.6%)	182 (67.4%)
mothers	96 (36.1%)	170 (63.9%)
sexuality	129 (44.2%)	163 (55.8%)

These texts are complex entities. Posts classified as conscious often match Taylor's '*ambiguous and rambling*' description, being statements of a situation which the poster seems to believe will benefit from interaction with people in the forum but which lacks a distinct statement of information need. Most of these posts describe situations rather than needs; they are providing an often complex description of their current situation as the basis for requesting help. Often posters are looking for people with similar experiences in the hope that those who have been through a similar situation can help with the right questions. The general sense of these posts is of someone who wishes to be talked through a situation to help structure the situation into one that can be made orderly and therefore solvable. This is often the task of professionals but here it is the informal lay community, and forum moderators who may be professionals, who is being asked to help based on the similarity of their experiences.

Formalized posts also often present complex situations and may contain multiple needs within one post but it is clear what the needs are and what a suitable answer may look like. The posts may be checking intuitions, e.g. asking if they are right to consider a situation as not normal, or checking their own calibrations, e.g. if they were right to be worried about an experience, and so can be describing complex situations. However, even if the descriptions are complex there is a clear and defined statement of need and a clear expectation of what will be the form of an answer.

3.3 Analysis

Our hypotheses are tested using the psycholinguistic Linguistic Inquiry and Word Count (LIWC) software [44], a dictionary-based toolkit for analysing text. LIWC contains dictionaries for various categories, e.g. positive emotions, cognitive words and perception words, and has been extensively as a means of analysing various properties of text in social media, interviews, and online text, e.g. [45-48], including recent work by Liu and Jansen who used a a simplified Chinese version of LIWC in a study to predict who is more likely to contribute to social Q&A sites [49] and Almatrafi et al. who used LIWC features to detect who most urgently needs help in MOOC forums [50]. In each section below, we explain which parts of LIWC we use to test our hypotheses.

Our data is not normally distributed so we use the non-parametric Mann Whitney independent samples test in our analysis. As we run a number of tests, we use conservative alpha value of 0.01 and as our hypothesis are directional we use a one-tailed test.

4. Findings

4.1 Hypothesis one

Our first hypothesis was that posts containing conscious information needs would be longer than those containing formalized needs. As shown in Table 2, this was the case for all four datasets and the differences in average word length per post was significant for all datasets. In all datasets the length is highly skewed [51]. The distribution within the conscious posts are more skewed indicating subsets of very long posts.

Table 2: Mean words per post for each dataset and skewness in parentheses words. *indicates a statistically significant difference.

	conscious	**formalized**	**p value**
diabetes	164.36 (2.37)	70.36 (1.53)	p<0.001*
finance	160.88 (2.62)	64.49 (2.12)	p<0.001*
mothers	205.49 (1.34)	94.72 (2.84)	p<0.001*
sexuality	186.60 (2.79)	83.99 (2.20)	p<0.001*

Posts containing conscious needs are longer than ones containing formalized ones for two reasons:

1. People who are unclear on what information they need are often unclear on what information to supply to obtain help. Most posts containing conscious needs are long because the poster does not know what information she needs; only that she has a situation that requires information. When such situations arise in an offline environment, in a doctor's surgery or a lawyer's office for example, we can create a dialogue where professionals use their insight and professional training to ask questions to help structure our information problem and move to a solution. In Internet forums such dialogues are possible, but not usually in real-time, and the tendency appears to be to provide as much information as possible as the poster does not know what will be useful information to those who may be able to help. In such posts, the poster themselves often give the indication that they realise they are providing a lot of information, some which may not be relevant, by the use of phrases such as *'Sorry this is a bit of an essay'*, '*and thank you if you've read this far*!' and '*don't know if it's relevant or not thanks*'.
2. Posters who are in highly emotional states may be less able to think through what information they need and so lack focus when expressing their needs. We know from everyday experience and studies such as [52] that emotion can interfere with cognition and so highly emotional states may lead to unfocussed posts. This has also been observed, as noted by Murphy [53], in library settings "*The patron's ability to communicate might also be facilitated or hindered by emotion*.".

To investigate this, in Table 3, we present the results of several Pearson correlation tests between post length (word count) and the percentage of positive and negative emotional words in each post. In all datasets, post length is positively and significantly correlated with a higher use of negative emotional words. If we focus on specific negative LIWC emotional categories – anger, anxiety and sadness – we see that greater use of these negative emotions is also positively and significantly correlated with post length. Taken together, these results indicate that posters who are in negative emotional states when posting are those who are more likely to be the ones who post longer, narrative posts rather than focussed posts describing formalized needs.

Table 3: Correlations between post length and percentage of emotional word use. *indicates statistically significant correlations.

	diabetes	**finance**	**mothers**	**sexuality**
positive emotional words	0.070 (p=0.208)	0.109 (p=0.073)	0.035 (p=0.566)	0.187* (p=0.001)
negative emotional words	0.278* (p<0.001)	0.478* (p<0.001)	0.366* (p<0.001)	0.217* (p<0.001)
anger words	0.321* (p<0.001)	0.341* (p<0.001)	0.527* (p<0.001)	0.375* (p<0.001)
anxiety words	0.329* (p<0.001)	0.334* (p<0.001)	0.369* (p<0.001)	0.332* (p<0.001)
sadness words	0.329* (p<0.001)	0.293* (p<0.001)	0.294* (p<0.001)	0.309* (p<0.001)

In summary, we can conclude that there is positive support for the first hypothesis and posts containing conscious information needs are longer than posts containing formalized information needs.

4.2 Hypothesis two

Our second hypothesis was that conscious needs are ones that reflect more uncomfortable emotional states and that the problematic situations, as neatly named by Cole [8], that contain conscious needs will be associated with negative emotional states. Specifically, we hypothesised that posts describing conscious information needs will contain an increased use of negative emotional words compared to posts describing formalized information needs.

To test this, we explored the LIWC categories *positive emotions* (including words like happy and good) and *negative emotions* (words such as hate and worthless), both of which are broad general categories of emotion, and specific emotions such as *anxiety* (words such as nervous, afraid, tense), *anger* (hate, kill) and *sad* (grief, cry, and sad). We also include the LIWC category *risk* (containing words such as danger and doubt) as risk can be perceived emotionally as well as cognitively.

In Table 4 we present the *rate* at which words from these LIWC categories are used, i.e. the percentage of words, on average, in each post that contains a word from each category. For example, in the diabetes dataset, on average, 2.52% of words in posts that contain formalized information needs are positive emotional words whereas only 2.31% of words in posts that contain conscious information needs are positive emotional words. We present the skewness in parentheses.

Table 4: Mean word rate per post for each dataset and skewness in parentheses. * indicates a statistically significant difference.

	conscious	**formalized**	**p value**
diabetes			
positive emotions	2.31% (0.84)	2.52% (1.34)	p=0.320
negative emotions	2.07% (1.47)	1.46% (1.78)	p<0.001*
anxiety	0.56% (2.75)	0.29% (3.37)	p<0.001*
anger	0.21% (4.56)	0.12% (5.24)	p=0.001*
risk	0.50% (1.63)	0.00% (2.98)	p=0.001*
sad	0.56% (2.80)	0.55% (4.29)	p=0.002*
finance			
positive emotions	2.16% (0.82)	2.96% (2.19)	p=0.393
negative emotions	1.33% (1.71)	0.56% (2.25)	p<0.001*
anxiety	0.48% (2.88)	0.10% (6.80)	p<0.001*
anger	0.16% (3.01)	0.06% (5.80)	p=0.001*
risk	0.34% (1.53)	0.00% (3.32)	p=0.001*
sad	0.30% (2.30)	0.21% (2.46)	p=0.001*
mothers			
positive emotions	2.35% (1.02)	2.61% (0.88)	p=0.539
negative emotions	2.63% (2.33)	1.71% (2.39)	p<0.001*
anxiety	0.75% (3.59)	0.54% (2.17)	p=0.001*
anger	0.41% (2.16)	0.28% (6.57)	p<0.001*
risk	0.44% (1.21)	0.25% (2.56)	p=0.001*
sad	0.66% (2.10)	0.39% (2.78)	p<0.001*
sexuality			
positive emotions	2.14% (0.91)	2.06% (9.13)	p=0.009*
negative emotions	2.59% (1.21)	2.00% (1.83)	p<0.001*
anxiety	0.86% (1.49)	0.64% (2.34)	p<0.001*
anger	0.29% (2.42)	0.15% (5.09)	p<0.001*
risk	0.99% (3.13)	0.94% (2.13)	p=0.001*
sad	0.48% (6.41)	0.36% (2.94)	p=0.003*

As can be seen in Table 4, for positive emotional words the tendency is for a non-significantly higher use of these words in posts containing formalized needs with the exception of the sexuality dataset where there is a significantly higher use of positive words in the posts containing conscious needs.

However, for negative emotions, there is a significantly higher use of this category of words in posts containing conscious needs in all four datasets. This indicates that people use a higher rate of negative emotions when creating posts that contain conscious information needs. If we focus on specific negative emotions, we see that for all four datasets there is a significantly higher use of anger, anxiety, risk and sad words in posts that contain conscious needs. Words from LIWC’s anxiety categories are more commonly expressed than words from the anger, risk or sad categories.

The mean frequency of these negative emotional words is low, often less than one word per hundred. However, these words often appear in phrases such as *'I’m quite anxious'*, ‘*I am now really worried something may be wrong with me*’, ‘*this just makes me really depressed'* and ‘*I'm feeling extremely depressed.*’ that cover the state of mind of the poster and therefore just one occurrence of words like *depressed* or *anxious* can be meaningful. The skewness for these categories tends to be higher for posts containing formalized needs. This indicates that these emotional words tend to be concentrated in fewer posts for formalized needs but more pervasive across posts containing conscious information needs.

Therefore, we can conclude that there is positive support for the second hypothesis and there is a higher rate of negative emotions expressed in posts containing conscious information needs over those posts containing formalized information needs.

4.3 Hypothesis three

Our third hypothesis was that in conscious needs, where we have not yet reached a full cognitive understanding of the information we need, we would rely more physical sensations as a source of understanding our situation and that posts describing conscious needs would rely more on sensory words. To test this, we used the LIWC categories *perceptual processes*, a general category reflecting perception and based on words such as *see*, *touch*, *listen*, and the specific categories *see* (words such as *view* and *saw*), *hear* (words such as *listen*) and *feel* (words such as *touch* and *felt*). The results are shown in Table 5.

Table 5: Mean word rate per post for each dataset and skewness in parentheses. * indicates a statistically significant difference.

	conscious	**formalized**	**p value**
diabetes			
perceptual	2.81% (3.57)	2.32% (3.53)	p=0.003*
see	0.69% (3.82)	0.61% (2.47)	p<0.001*
hear	0.43% (2.16)	0.26% (3.23)	p<0.001*
feel	1.08% (4.26)	0.90% (6.37)	p=0.002*
finance			
perceptual	1.11% (4.73)	0.96% (5.41)	p=0.017
see	0.35% (3.03)	0.40% (8.86)	p=0.016
hear	0.40% (1.80)	0.44% (3.00)	p=0.024
feel	0.23% (4.25)	0.07% (3.82)	p<0.001*
mothers			
perceptual	1.99% (0.87)	1.78% (2.01)	p=0.006*
see	0.60% (1.99)	0.59% (2.46)	p=0.001*
hear	0.36% (2.51)	0.35% (3.18)	p=0.007*
feel	0.92% (1.52)	0.80% (3.91)	p=0.001*
sexuality			
perceptual	2.63% (1.41)	2.40% (2.09)	p=0.016
see	0.61% (2.67)	0.56% (2.70)	p=0.001*
hear	0.58% (5.93)	0.42% (3.47)	p<0.001*
feel	1.21% (2.67)	1.13% (2.26)	p=0.006*

From Table 5, we see that the evidence is generally supportive of the hypothesis. The general perception category of words was significantly higher in posts containing conscious needs in only two datasets but the see/hear words were used at a significantly higher rate in three out of four datasets and feel words used at a significantly higher rate in all datasets.

Posters use these sensory words in a mixture of senses. The word ‘see’ for example being used to reflect interactions (‘…*see the baby*…’, ‘…*see all my friends*…’, ‘…*see my GP*…’), visual perception (‘…*see he enjoys it*…’), deduction and discernment (‘…*see how it effected me*…’, ‘…*to see if the antibiotics would stop*…’), cause (‘…*seeing as I couldn’t eat*…’) and imagined situations (‘…*we both see him as the dad*…’). Feeling can also be used in various senses, reflecting physical sensations (‘…*the midwife was feeling my belly*…’, ‘…*feeling too hot*…’, ‘…*but if i dont i feel sick*…’), experiencing sensations or emotions (‘…*I feel really abnormal*…’, ‘…*I feel like I'm still being punished*…’, ‘…*I feel so alone right* now…’) and opinions (‘…*hey feel that my unborn baby*…’, ‘…*I feel it would be best to*…’).

Reading across the posts, there is no qualitative difference between the uses of various senses as both formalized and conscious posts contain examples of all these sense. Rather, the difference is that conscious posts contain a *higher rate of use* of sensory words, reflecting more attention to physicality. As with the emotional words, the skewness for these categories tends to be higher for posts containing formalized needs, indicating these words are more pervasive across posts containing conscious information needs.

These findings provide support our third hypothesis that, we rely more on sensory signals, or words that express senses, when in the early stages of information need development.

4.4 Hypothesis four

Our final hypothesis was that posts containing conscious needs would reveal more words relating to active cognition as the poster is trying to understand a problematic situation. We used several LIWC categories to investigate this. Firstly, the general *cognitive processes* category that reflects domain-general cognitive words (such as *cause*, *ought*) and then secondly specialist categories that reflect different types of cognition: *insight* (based on words such as *think*, *know*, *consider*), *causation* (based on words such as *because*, *effect*), *discrepancy* (words such as *should* and *would*), *tentative* (*maybe*, *perhaps*), *certainty* (*always*, *never*) and *differentiation* (*hasn't*, *but*, *else*) to determine whether early and late stage information needs are thought about differently.

We also include three temporal categories *focus past*, *focus present* and *focus future* that measure the use of past/present/future tense words and references to past/present/future events. We include these to see if there are any differences in the time periods being discussed; are some posts more focussed on unchangeable past events, and therefore possibly still trying to come to terms with them, or looking forward to possible futures? Table 6 presents the results for the cognition categories and Table 7 for the temporal categories.

Table 6: Mean word rate per post for each dataset and skewness in parentheses. * indicates a statistically significant difference.

	conscious	**formalized**	**p value**
diabetes			
cognitive	13.41% (-0.01)	13.94% (0.61)	p=0.258
insight	3.16% (1.11)	2.87% (0.84)	p=0.004*
causation	1.64% (0.66)	2.03% (1.78)	p=0.037
discrepancy	1.55% (1.12)	1.39% (1.82)	p=0.273
tentative	3.84% (0.90)	3.43% (1.46)	p=0.032
certainty	1.37% (1.63)	1.12% (1.82)	p=0.016
differentiation	3.84% (0.56)	3.89% (0.99)	p=0.099
finance			
cognitive	13.55% (1.63)	12.86% (0.65)	p=0.201
insight	2.34% (2.80)	2.28% (1.53)	p=0.110
causation	1.74% (2.41)	1.79% (1.99)	p=0.055
discrepancy	2.36% (0.58)	2.14% (1.31)	p=0.386
tentative	3.79% (0.41)	3.49% (2.06)	p=0.415
certainty	1.22% (1.01)	1.02% (2.33)	p<0.001*
differentiation	4.05% (1.12)	4.04% (1.90)	p=0.056
mothers			
cognitive	12.92% (1.47)	12.52% (0.47)	p = 0.089
insight	2.39% (1.05)	2.28% (1.49)	p = 0.094
causation	1.37% (5.90)	1.15% (1.27)	p = 0.026
discrepancy	2.05% (0.74)	1.83% (1.14)	p=0.493
tentative	3.27% (1.44)	2.92% (2.21)	p=0.001*
certainty	1.53% (0.95)	1.33% (1.46)	p<0.001*
differentiation	4.17% (0.88)	3.85% (0.50)	p=0.029
sexuality			
cognitive	15.26% (0.31)	13.95% (0.42)	p=0.005*
insight	2.71% (0.58)	2.18% (1.34)	p<0.001*
causation	1.61% (0.89)	2.15% (2.59)	p=0.138
discrepancy	2.41% (1.35)	2.09% (1.43)	p=0.009*
tentative	3.86% (0.53)	4.09% (1.03)	p=0.442
certainty	1.60% (1.06)	1.15% (1.23)	p<0.001*
differentiation	5.17% (0.13)	4.48% (0.77)	p=0.003*

The first conclusion from Table 6 is that cognition is strong within both classes of posts. The frequencies of word use are higher than the categories investigated in sections 4.1-4.3 with the post frequency of general cognitive words at about 13-14% of total words in the posts. For only the sexuality dataset is there a significant difference in the rate of word use for the general cognitive processes category indicating that general cognition is as strong in each type of post. In three out of the four datasets word relating to *certainty* are significantly more common in posts containing conscious information needs and for two out of the four datasets is there are a higher use of words from the *insight* category of words. Tausczik and Pennebaker suggests that this category of words are reflective of people trying to actively process or reappraise an event or situation, [54].

Beyond these two differences, and the observation that for the sexuality dataset, words from most of the cognitive categories are used more often in posts containing conscious information needs, there is no general evidence that people are using more/fewer cognitive words or using different cognitive words when describing conscious compared to formalised information needs. That is, we do not have solid evidence that people think differently at earlier stages of information need development but do have evidence that they think a lot based on the frequency of words from these LIWC categories expressed in these posts. There are no solid patterns regarding the skewness of word use but the skewness scores are notably lower than emotion and perception results indicating, again, that cognition is widely used across posts.

If we look at the temporal categories, Table 7, we see that the present tense is commonly used across both categories and datasets. As we noted in section 3.1, the value of online postings is that they are data written by people who are actively thinking about current situations. This result, demonstrating the high level of words about the present validates this claim. The low skewness scores for most categories and datasets indicates that time is pervasive to discussions of the situation being presented.

Table 7: Mean word rate per post for each dataset and skewness in parentheses. * indicates a statistically significant difference.

	conscious	**formalized**	**p value**
diabetes			
past	5.13% (0.84)	4.20% (0.60)	p=0.005*
present	12.56% (0.67)	12.91% (0.50)	p=0.382
future	0.99% (3.09)	1.19% (2.30)	p=0.145
finance			
past	3.21% (0.62)	2.92% (1.27)	p=0.022
present	13.62% (0.70)	14.75% (1.03)	p=0.076
future	1.33% (0.87)	1.37% (1.70)	p=0.033
mothers			
past	4.32% (0.74)	2.86% (0.87)	p<0.001*
present	14.58% (-0.07)	14.29% (-0.20)	p=0.249
future	1.54% (2.05)	1.16% (1.95)	p<0.001*
sexuality			
past	6.22% (0.40)	5.41% (1.62)	p=0.008*
present	13.06% (1.03)	14.19% (1.68)	p=0.056
future	1.07% (0.41)	1.07% (0.73)	p=0.113

In three out of four datasets, there is a significantly higher use of past focus words in posts containing conscious needs than those containing formalized needs. Past focus words are typically being used in the posts to describe a situation that has occurred as a means of explaining why information is required.

Cognition is obviously strong in the situations that encouraged these people to post online. Some types of cognition, such as insight, are stronger in two datasets whereas most others are not. The support for hypothesis four is therefore weak and this hypothesis is not supported based on our current evidence. However, the findings regarding temporal dimensions suggest that there may be different types of cognition being expressed in these posts and it is worthy of future research to examine this in more detail.

5. Limitations

This work has used forum posts as a source of data to investigate linguistic differences between posts containing early and later stage information needs. We specifically focussed on the comparison between Taylor's conscious and visceral information needs. Information needs, as conceptualised by Taylor, are a spectrum so there is not clear, single characteristic to define when a need is at one level or another. Our approach to classification, section 3.2, is simple but seems reliable from our inter-coder test. Working only from textual descriptions of need does mean that we cannot compare our classification of need level against the owner of that need. However, working with online posts does provide far more data than methods such as interviews and allows us to see patterns that we perhaps could not see using other investigative methods which work with far smaller datasets. We deliberately chose forums representing different areas of life to be able to generalise more across information needs. However, these are all one type of data and it would be worth contrasting results from this type of dataset to results obtained from other types of data.

6. Discussion

We first describe our findings in relation to research on information needs then the implication for those who host online Q&A systems and discussion forums.

Information needs theory

Information need is a broad term that is used across disciplines but often used inconsistently with the term variably describing the information that is needed, the situation that requires information or even just the recognition that a person's current cognitive state is somehow insufficient. Various contributions, e.g. [13], have classified approaches to understanding information needs or what affects information needs highlighting factors such as environmental or demographics factors and other authors have contributed to our understanding of the motivational factors that give rise to information needs, e.g. [55, 56]. However, the language of information needs still remains loose with a lack of clarity around how needs differ except when used to discuss the topic of the information need (financial information needs, health information needs, etc.). Indeed, information needs are more commonly classified by topic than other attributes of the need such as complexity of need even if such attributes are important in determining information seeking success: we know for example from studies such as [57, 58] that how people interact at the early stages of a search differ from the later stages.

In this paper, inspired by Taylor's classification of needs, we examined the linguistic differences between statements of need that either contained a conscious or formalized information need statement. We used over 1100 Internet forum posts as written descriptions of a problematic situation or expressed need, with the advantage that we could examine the needs as they were described by the person with the need.

Our study was guided by four hypotheses derived from the literature on information needs. Our first hypothesis was that posts containing conscious needs would be longer than posts containing formalized needs. This was conclusively proven with those posts that are at the level of conscious needs being significantly longer than those posts at the level of formalized needs. Conscious needs in our datasets are typically presented as situations that are troubling to the person posting. This finding supports the arguments of Belkin and others summarised in section 2.1, that situations can be an important method of presenting needs when one cannot form questions or statements of information requirements.

These situations may have an emerging focus which is in the process of developing into formalized needs, or they may have no focus requiring an outsider to structure the situation into questions that can be answered, a plan for action or a direction to where else may be a good source of support. Examining the differences between these two types of post we believe can be a fruitful way of understanding how emerging information needs develop and why some early stage information needs do not develop further.

Both Genuis and Bronstein [38] and Ruthven et al. [43] have demonstrated that many people used online forums to make sense about what is 'normal'. In some of the posts we analysed, posters were asking about normality, e.g. is it normal for a child to behave in a particular way or are my blood levels normal; in other posts, posters were asking about how to think about a situation (e.g. was I deceived?). Many of these were in the conscious level posts leading interesting direction of future research to examine how people understand whether they are in a need situation at all.

Our second hypothesis was that posts containing conscious needs reflect more uncomfortable emotional states and that the language used in the posts will describe these emotional states. Many writers have discussed how a lack of information has negative consequences in terms of anxiety or worry. Conscious needs, in Taylor's categorisation, reflect an '*ill-defined'* part of someone's life, a situation where uncertainty may be high, especially uncertainty about what may be required to move forward. Therefore, we predicted that when uncertainty is high, the negative emotions we feel will come out in our descriptions of the situations. This was generally true, particularly for anxiety-related words. That anxiety was a dominant emotion expressed also fits with the findings of Kuhlthau, Brashers and others who note the importance of this emotional/physical complement to the cognitive manifestation of uncertainty [33, 36].

That the presence of conscious needs within posts correlated with the presence of negative emotions suggests that strong emotions can be indicative of early stage information needs. This may be, as we suggest in section 4.1, that emotions interfere with the ability to focus sufficiently in order to create formal statements of need or it may be that the early stages of needs are ones that more emotional and these negative emotional states lead to information seeking [59]. Both viewpoints are supported by literature, e.g. [41], and both may be factors at this early stage of information need development: problematic situations that give rise to negative emotions states compel us to find solutions [24, 25, 59], but the negative emotional states may make it difficult for us to think about what information we need.

If we are struggling to understand our situation, and therefore what information may help, then we may have to rely more on sensory signals as a source of understanding our situation when unable to describe our information needs. This led to the hypothesis that posts describing conscious information needs would contain more words relating to sensory perceptions. We found some evidence to support this hypothesis, particular on the use of 'feeling' words. Perception is generally under-studied in information seeking with far more attention given to cognitive aspects of information needs. However, as noted in section 2.3, other fields recognise the importance of physicality as a source of understanding of the world, particularly when something may be going wrong and things 'feel' not right. Our findings indicate that linguistic expression of early stage information needs do seem to involve more use of words describing physical sensation. This fits with Godbol's observation of '*sensations as informative elements*' used to understand situations – with the emphasis on situations not needs [28, 30]. More work certainly needs done here, especially as LIWC only allows us to investigate certain types of sensory words but our evidence would indicate that it is an area worth pursuing.

Our final hypothesis was that early stage information needs might display more active cognition as people tried to understand their situation and that posts containing conscious information needs would express more cognition words. Cognition was very strong in our posts but there was no substantial difference in the use of cognition words between posts containing conscious and formalized needs.

A clearer pattern was that, in posts describing conscious needs we see more focus on the past and more references to what has happened. Even though situations may be ambiguous or uncertain, describing this situation may be the easiest way to ask for information as we know what has happened to create a problematic situation. That is, the situation may be the one thing we are most confident about, even if we are unsure about how to act within the situation.

Taking these findings together we see that conscious and formalized needs, which in Taylor's characterisation of needs reflect different psychological states, reveal themselves in these online posts by a differing use of words. Posts containing conscious information needs, earlier stage information needs, are more emotional in language, are more based on sensory properties and more narrative in form with a stronger focus on the past. We suggest, therefore, that when posters cannot ask directly for what information they want, i.e. cannot present a formalized need, they instead describe what they can talk about: their emotions, their sensations, and what has happened to lead them to seek help.

Practical implications

In the following two sections I present some of the implications for this work for those moderating online discussion forums and those developing automatic techniques for classifying posts to forums.

Good and bad questions: answer success and failure

A core issue in the literature on social Q&A interactions and online information seeking is the idea that some information requests are better than others [60-63] with good questions seen as those which are more likely to receive an answer [60]. Choi et al. [60], looking at factual questions on Yahoo Answers!, found that textual features, such as the level of clarity in a question, can be important in predicting if a question will receive an answer of not. Chua et al. [62] also found that level of details, specificity, clarity were important determiners in the likelihood of questions receiving answers and Shah et al. found that providing too much information or providing too little information in a request could both be reasons for answer failure on social Q&A sites [61]. Our findings propose a differentiation that may be useful: that some postings are reflecting information needs that are less well developed and therefore may be more difficult to answer without interaction.

Some features from Chua et al.'s study (such specificity and clarity) would seem to be more indicative of later stage, formalized information needs and whereas longer posts that provide more information is characteristic of postings containing early stage, conscious level information needs. Therefore 'good' questions may be ones that are reflecting later stage information needs and 'bad' questions ones that contain earlier stage information needs.

Similar to the suggestion made by Kitzie et al. [64] our results could be useful in helping posters create good requests in the first place by linguistically analysing their posts as they are being written. This may allow the system to suggest better ways to frame questions. If posters cannot provide focussed questions then linguistic analyses can help forum moderators recognize that the posts do not contain developed questions and may require interaction. This then may particularly help those searchers who are engaged in more exploratory types of information interactions.

Moderation and automatic classification

Several authors have examined the various types of questions and motivations for asking questions in online environments. Zhang, for instance, detailed types of question goals, e.g. understanding, verifying (yes/no questions), fact-finding, seeking practical advice, seeking personal experiences, or seeking recommendations, and affective goals for interacting online, e.g. reducing uncertainty, clearing suspicions, avoiding embarrassment [65]. Westbrook and Zhang also noted different types of questions and answers on cervical cancer forum including facts, explanations, stories and emotional support and claimed that '*Posters expect to receive very personalized responses to their* *requests*.' [66]. Shah et al. proposed different types of questions on social Q&A sites: factual, advice, opinion seeking and social questions and demonstrated that adding information on question type could improve the performance of automatic classifiers based on textual features [63]. The implications from much of this work are that systems may wish to understand what is the real need faced by the poster and tackle that. Our work contributes to this goal by showing that we can classify posts both into level of information need contained but also, section 4.2, the emotional state of the person who is posting. Thereby forum moderators and participants gain additional information on how to answer online postings.

Forum moderation can prevent inappropriate responses but can also result in the forum being a useful archive for future users [67]. In some studies, e.g. [67], posts that do not contain enough information can be received critically. In our study we have shown that people with early stage information needs often produce too much information because they do not know what information may be useful to those trying to help. Therefore studies such as ours could help moderators recognise situations when people are struggling to provide good requests. Huh et al. showed that linguistic analyses can be useful for determining automatically which posts needed attending from a moderator and which did not, [68]; our findings can be used to determine at what level is an information need expressed in an online post and therefore *how* moderators should respond.

7. Conclusion

This paper has examined one of the most famous contributions to Information Science, Taylor's conceptualisation of information needs, from the novel perspective of analysing the language used when expressing need. Asking for information online is now an everyday activity. In doing so, posters are creating large repositories of textually described needs that can be investigated to provide large-scale analyses of information needs. Here, we use over 1100 posted need statements to analyse the language used at different levels of information need development, demonstrating that conscious and formalized information needs are different in the language that they contain.

Specifically, we show that

- descriptions of early stage information needs are longer and more narrative than those describing later stage information needs;
- descriptions of early stage information needs contain a higher rate of negative emotional words than those describing later stage information needs;
- descriptions of early stage information needs contain a higher rate of certain sensory words than those describing later stage information needs.
- linguistic features can differentiate between conscious and formalized needs. This important contribution can facilitate new research into how these needs differ and how we can detect and support different kinds of search activity.

Future research is needed to break down these results in more detail to consider more precisely which words are most powerful in making these distinctions and also to test these hypotheses on different datasets on other areas of life. However, we hope to provide a new focus on one of the most significant contributions to Information Science, and a new way of theorising about information needs.

Acknowledgements

I am very grateful to the two anonymous reviewers and Dr Luanne Freund, Associate Editor, for their very insightful and constructive comments on this paper and to my colleagues in Strathclyde's Information School for their input into the early stages of this work.

References

1. Taylor, R.S., *Question-negotiation and information seeking in libraries.* College & research libraries, 1968. **29**(3): p. 178-194.
2. Cole, C., *Information Need and the Beginning of Information Search*, in *Encyclopedia of Information Science and Technology, Third Edition*, D.B.A. Mehdi Khosrow-Pour, Editor. 2015, IGI Global: Hershey, PA, USA. p. 4117-4128.
3. Tyckoson, D.A., *Question-negotiation and information seeking in libraries: a timeless topic in a timeless article.* College & Research Libraries, 2015. **76**(3): p. 247-250.
4. Hoenkamp, E.C., *About the 'Compromised Information Need' and Optimal Interaction as Quality Measure for Search Interfaces*, in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2015, ACM: Santiago, Chile. p. 835-838.
5. Pálsdóttir, Á. *From Noticing to Suspecting: The Initial Stages in the Information Behaviour of Informal Caregivers of People with Dementia*. in *International Conference on Human Aspects of IT for the Aged Population*. 2017. Springer.
6. Jansen, M. and S. Hoppenbrouwers, *Framework for executing, measuring and optimizing the sales process*. 2017, Radboud University.
7. Chang, Y.-W., *The influence of Taylor's paper, question-negotiation and information-seeking in libraries.* Information Processing & Management, 2013. **49**(5): p. 983-994.
8. Cole, C., *A theory of information need for information retrieval that connects information to knowledge.* Journal of the American Society for Information Science and Technology, 2011. **62**(7): p. 1216-1231.
9. Lundin, J., *Towards a normative conceptual framework for information-seeking studies in technical communication*, in *Proceedings of the International Conference on Information Systems and Design of Communication*. 2014, ACM: Lisbon, Portugal. p. 15-19.
10. Agarwal, N.K., *Towards a definition of serendipity in information behaviour.* Information Research: An International Electronic Journal, 2015. **20**(3): p. n3.

11. Timmins, F., *Exploring the concept of 'information need'.* International journal of nursing practice, 2006. **12**(6): p. 375-381.

12. Savolainen, R., *Information need as trigger and driver of information seeking: a conceptual analysis.* Aslib Journal of Information Management, 2017. **69**(1).

13. Ormandy, P., *Defining information need in health–assimilating complex theories derived from information science.* Health expectations, 2011. **14**(1): p. 92-104.

14. Lundh, A., *Studying information needs as question-negotiations in an educational context: a methodological comment.* Information Research: An International Electronic Journal, 2010. **15**(4): p. n4.

15. Belkin, N.J., R.N. Oddy, and H.M. Brooks, *ASK for information retrieval: Part I. Background and theory.* Journal of documentation, 1982. **38**(2): p. 61-71.

16. Savolainen, R., *Conceptualizing information need in context.* Information Research, 2012. **17**(4).

17. Dervin, B., *Sense-making theory and practice: an overview of user interests in knowledge seeking and use.* Journal of knowledge management, 1998. **2**(2): p. 36-46.

18. Savolainen, R., *Elaborating the motivational attributes of information need and uncertainty.* Information Research, 2012. **17**(2).

19. Anderson, T.D., *Uncertainty in Action: Observing Information Seeking within the Creative Processes of Scholarly Research.* Information Research: an international electronic journal, 2006. **12**(1): p. n1.

20. Heinstrom, J., *From fear to flow: personality and information interaction*. 2010: Elsevier.

21. Lopatovska, I. and I. Arapakis, *Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction.* Information Processing & Management, 2011. **47**(4): p. 575-592.

22. Kuhlthau, C.C., *Inside the search process: Information seeking from the user's perspective.* Journal of the American Society for information Science, 1991. **42**(5): p. 361.

23. Brashers, D.E., D.J. Goldsmith, and E. Hsieh, *Information seeking and avoiding in health contexts.* Human Communication Research, 2002. **28**(2): p. 258-271.

24. Nahl, D. and D. Bilal, *Information and emotion: The emergent affective paradigm in information behavior research and theory*. 2007: Information Today, Inc.

25. Zhang, Y. *Contextualizing consumer health information searching: an analysis of questions in a social Q&A community*. in *Proceedings of the 1st ACM International Health Informatics Symposium*. 2010. ACM.

26. Lloyd, A., *Information literacy landscapes: an emerging picture.* Journal of Documentation, 2006. **62**(5): p. 570-583.

27. Olsson, M. and A. Lloyd, *Being in place: embodied information practices.* Information Research, 2017. **22**(1).

28. Yates, C., *Exploring variation in the ways of experiencing health information literacy: A phenomenographic study.* Library & Information Science Research, 2015. **37**(3): p. 220-227.

29. Brashers, D.E., et al., *Communication in the management of uncertainty: The case of persons living with HIV or AIDS.* Communication Monographs, 2000. **67**(1): p. 63-84.

30. Godbold, N., *Listening to bodies and watching machines: Developing health information skills, tools and services for people living with chronic kidney disease.* Australian Academic & Research Libraries, 2013. **44**(1): p. 14-28.

31. Douw, G., et al., *Nurses' worry or concern and early recognition of deteriorating patients on general wards in acute care hospitals: a systematic review.* Critical Care, 2015. **19**(1): p. 230.

32. King, L. and J.M. Clark, *Intuition and the development of expertise in surgical ward and intensive care nurses.* Journal of advanced nursing, 2002. **37**(4): p. 322-329.

33. Brashers, D.E., *Communication and uncertainty management.* Journal of Communication, 2001. **51**(3): p. 477-497.

34. Babrow, A.S., *Communication and problematic integration: Milan Kundera's "lost letters" in the book of laughter and forgetting.* Communications Monographs, 1995. **62**(4): p. 283-300.

35. Ingwersen, P., *Search procedures in the library—analysed from the cognitive point of view.* Journal of documentation, 1982. **38**(3): p. 165-191.

36. Kuhlthau, C.C., *Seeking meaning.* Norwood, NJ: Ablex, 1993.

37. Bronstein, J., *Is this OCD?: Exploring conditions of information poverty in online support groups dealing with obsessive compulsive disorder.* Information Research, 2014. **19**(4).

38. Genuis, S.K. and J. Bronstein, *Looking for "normal": Sense making in the context of health disruption.* Journal of the Association for Information Science and Technology, 2016: p. n/a-n/a.

39. Hasler, L., I. Ruthven, and S. Buchanan, *Using internet groups in situations of information poverty: Topics and information needs.* Journal of the Association for Information Science and Technology, 2014. **65**(1): p. 25-36.

40. Eysenbach, G. and J.E. Till, *Ethical issues in qualitative research on internet communities.* BMJ, 2001. **323**(7321): p. 1103-1105.

41. Ruthven, I., Buchanan, S., & Jardine, C., *Isolated, overwhelmed and worried: young first-time mothers asking for information and support online.* Journal of the Association for Information Science and Technology, 2018. **in press**.

42. Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data.* biometrics, 1977: p. 159-174.

43. Ruthven, I., Buchanan, S., & Jardine, C., *Relationships, environment, health and development: the information needs expressed online by young first-time mothers.* Journal of the Association for Information Science and Technology, 2018. **in press**.

44. Pennebaker, J.W., et al., *The development and psychometric properties of LIWC2015*. 2015: https://repositories.lib.utexas.edu/handle/2152/31333.

45. Chowdhury, M.F.M., et al. *Fbk: Sentiment analysis in twitter with tweetsted.* in *Second Joint Conference on Lexical and Computational Semantics (*

SEM): Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval'13). 2013.

46. Harman, G.C.M.D.C., *Quantifying mental health signals in Twitter.* ACL 2014, 2014. **51**.

47. Savage, D.A. and B. Torgler, *The emergence of emotions and religious sentiments during the September 11 disaster.* Motivation and Emotion, 2013. **37**(3): p. 586-599.

48. Yao, Y. and S. Yarosh. *Group Finder: Finding the" Right": Online Support Groups for People in Recovery*. in *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. 2016. ACM.

49. Liu, Z. and B.J. Jansen, *Identifying and predicting the desire to help in social question and answering.* Information Processing & Management, 2017. **53**(2): p. 490-504.

50. Almatrafi, O., A. Johri, and H. Rangwala, *Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums.* Computers & Education, 2018. **118**: p. 1-9.

51. Bulmer, M.G., *Principles of statistics*. 1979: Courier Corporation.

52. Kuppens, P., D. Champagne, and F. Tuerlinckx, *The Dynamic Interplay between Appraisal and Core Affect in Daily Life.* Frontiers in Psychology, 2012. **3**: p. 380.

53. Murphy, S.A., *The reference narrative.* Reference & User Services Quarterly, 2005: p. 247-252.

54. Tausczik, Y.R. and J.W. Pennebaker, *The psychological meaning of words: LIWC and computerized text analysis methods.* Journal of language and social psychology, 2010. **29**(1): p. 24-54.

55. Savolainen, R., *Approaching the motivators for information seeking: The viewpoint of attribution theories.* Library & Information Science Research, 2013. **35**(1): p. 63-68.

56. Savolainen, R., *Emotions as motivators for information seeking: A conceptual analysis.* Library & Information Science Research, 2014. **36**(1): p. 59-65.

57. Moshfeghi, Y. and J.M. Jose, *On cognition, emotion, and interaction aspects of search tasks with different search intentions*, in *Proceedings of the 22nd international conference on World Wide Web*. 2013, ACM: Rio de Janeiro, Brazil. p. 931-942.

58. White, R.W., I. Ruthven, and J.M. Jose, *A study of factors affecting the utility of implicit relevance feedback*, in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. 2005, ACM: Salvador, Brazil. p. 35-42.

59. Savolainen, R., *Approaching the affective factors of information seeking: the viewpoint of the information search process model.* Information Research: An International Electronic Journal, 2015. **20**(1): p. n1.

60. Choi, E., V. Kitzie, and C. Shah, *A machine learning-based approach to predicting success of questions on social question-answering*, in *iConference 2013* 2013. p. 409-421.

61. Shah, C., et al., *"How much change do you get from 40$?"–Analyzing and addressing failed questions on social Q&A.* Proceedings of the Association for Information Science and Technology, 2012. **49**(1): p. 1-10.

62. Chua, A.Y. and S. Banerjee, *Answers or no answers: Studying question answerability in Stack Overflow.* Journal of Information Science, 2015. **41**(5): p. 720-731.

63. Shah, C., V. Kitzie, and E. Choi. *Questioning the Question--Addressing the Answerability of Questions in Community Question-Answering*. in *System Sciences (HICSS), 2014 47th Hawaii International Conference on*. 2014. IEEE.

64. Kitzie, V., E. Choi, and C. Shah, *From bad to good: An investigation of question quality and transformation.* Proceedings of the American Society for Information Science and Technology, 2013. **50**(1): p. 1-4.

65. Zhang, Y., *Toward a layered model of context for health information searching: An analysis of consumer-generated questions.* Journal of the American Society for Information Science and Technology, 2013. **64**(6): p. 1158-1172.

66. Westbrook, L., *"I'm Not a Social Worker": An Information Service Model for Working with Patrons in Crisis.* The Library, 2015. **85**(1).

67. Bullard, J., *It takes a jerk to make a conversation into an archive.* 2013.

68. Huh, J., M. Yetisgen-Yildiz, and W. Pratt, *Text classification for assisting moderators in online health communities.* Journal of biomedical informatics, 2013. **46**(6): p. 998-1005.