Adjectives, communities, and taxonomies of evaluative meaning

Neil Millar and Susan Hunston

University of Tsukuba / University of Birmingham

From a corpus consisting of free text comments submitted to the website RateMyProfessors.com, adjectives describing people have been identified and clustered using Principal Components Analysis. This follows the Meaning Extraction method used by Chung & Pennebaker (2007). The outcome is a set of seven factors, each of which is interpretable as representing a dimension of meaning along which individual professors have been evaluated. These dimensions are in turn discussed using Martin & White's (2005) parameters of Judgement and Appreciation, and Coffin's (2005) concept of evaluative voices. It is argued that contributors to RateMyProfessors.com have available three distinct voices: 'novice intellectual', 'consumer', and 'subordinate'. The paper demonstrates how a bottom-up, statistical technique may be used to provide the initial data for identifying evaluative parameters. It raises the possibility that such parameters may be specific to individual discourse communities. It therefore offers a complementary and problematizing alternative to top-down, researcher-driven taxonomies of evaluative meaning.

1. Introduction

This paper asks to what extent a statistical approach to adjective co-occurrence can illuminate the identification of categories of Judgement (Martin 2000; Martin & White 2005) in a corpus of short texts whose purpose is to evaluate the performance of teaching staff in universities. It presents the results of a quantitative study of evaluative adjectives in a web-based collection of reviews of university professors (*RateMyProfessors.com*). The groupings of adjectives derived from this method are used to problematise the concept of taxonomies of evaluative meaning and to examine taxonomies that might emerge from a virtual, on-line community. In doing so, the paper explores the consequences of deriving a taxonomy that is corpus-specific, based on form and on associative frequency — essentially, a bottom-up, quantitative methodology — rather than on an analyst-driven, interpretative and top-down methodology. The study is used to support an argument that taxonomies of meaning may be context-dependent and unique to the communities that generate them.

2. Background

The website used in this study, *RateMyProfessors.com* (*RMP*), has ideological significance as well as practical usefulness for the researcher. It is reportedly relied upon extensively by students in the US in selecting universities and courses (Davidson & Price 2009; Brown *et al*. 2009), but it has also been criticised for encouraging contributors to assess instructors in terms of personality, or even appearance, rather than expertise (Felton *et al*. 2004, 2008). Discussions of this site are relevant to more general debates about forms of student evaluation of their teachers, which gain importance as those evaluations increasingly feed into employer assessments of teacher competence (e.g. Chen & Hoshower 2003; Marsh 2007; Shevlin *et al*. 2000). As noted below ('Data Selection'), contributors to the website can give their selected professor a series of numerical scores, and can also add free text content. We draw on these free texts in this study.

The texts posted on *RMP* may be considered in terms of theories of context, affinity space, and community. Within Systemic Functional Linguistics, context is theorised as comprising the variables of Field, Tenor and Mode. In the *RMP* texts, evaluation or appraisal of the target individuals is key to the Field of discourse — the texts exist to offer opinions on the efficacy of teaching staff. The Mode is constitutive, and the channel is electronic, that is, written but with many of the informal features associated with spoken language. In terms of Tenor, the website contributors are peers advising each other, but there is very little interaction: writers rarely respond to each other's texts, and are not necessarily acquainted with one another. This raises the question of the extent to which contributors to the site genuinely constitute a social group within which a genre may be said to exist. Gee (2005) has suggested the concept of 'affinity space' to account for groups of individuals who are co-located in an identifiable physical or virtual space but who have insufficient mutual sense of 'belonging' or common purpose to support the notion of a 'community of practice'. He gives the example of a computer game (*Age of Mythology*) as an example of individuals interacting with a body of content and with each other, in ways that are determined by the way the game has been written. It could be argued that *RMP* similarly brings together individuals who are not known to each other and who may have different motives for contributing to the site and whose only common factor is that they do write free text comments for *RMP*.

For our purposes, however, the concept of 'community' to describe the contributors to *RMP* is a useful one. It draws attention to an implicit set of shared values which we wish to exploit in examining the language of the texts. The relationship between evaluative values and community, with Appraisal as a template for investigation, is an emerging theme in research into on-line communication (Drasovean & Tagg subm.; Knight 2008; Zappavigna 2011). These researchers investigate 'virtual' communities, that is, communities that exist only by means of participation in on-line exchange of views. Knight studies a corpus of texts transmitted using MSN Messenger®. Zappavigna focuses on tweets using the hash-tag 'Obama'. Drasovean & Tagg investigate contributions to a TED discussion forum. Citing Stenglin (2004), Knight (2008: 484) expresses the connection between community and evaluation thus: "interactants construct social identities by expressing attitudes towards experiences that are part of a shared community". Zappavigna and Drasovean & Tagg concur, arguing that common ground in the community is established by a uniformity of practice in expressing attitude. This is particularly important in contexts of new media, where community membership is transient or 'ephemeral' (Drasovean & Tagg subm.: 10) and not defined by either geographical space or even by consistency of participation in 'conversations', but by a (possibly temporary) shared interest in a topic and shared presuppositions drawn on in making evaluations. The community that we investigate in this study is comparable to those in the above studies: it exists only in that its members have contributed one or more comments to the website *RateMyProfessors*. Members do not necessarily know each other and (unlike the participants whose texts are studied by Knight, or Drasovean & Tagg) do not usually respond to each other's contributions.

This research into Appraisal and community might be seen as a continuation of a longer tradition that identifies similarities and differences in the use of evaluative or interpersonal resources, for example in texts produced by discourse communities aligned with particular academic disciplines or viewpoints. In a number of studies, Hyland (2005, 2009: 67–95, 2012: 178–193), Charles (2006, 2007), and Hiltunen (2010), among others, find significant differences between academic disciplines in terms of the interactive or metadiscoursal resources used. Most significantly for our work, Coffin (2002), following Iedema *et al*. (1994), identifies consistent configurations of Appraisal resources to distinguish between dominant 'voices' in texts written by students writing essays in History. According to Coffin (2002: 520), the 'Reporter', 'Interpreter' and 'Adjudicator' voices are each associated with various probabilities of types of Judgement and Appreciation. For example, both Social Sanction and Social Esteem (the main categories of Judgement) are rare in the Reporter voice but likely to be present in texts expressing the Interpreter or Adjudicator voices.

The role of evaluation in community construction is not surprising. Early research on narrative (e.g. Polanyi 1985; Schiffrin 1984; and more recently Eggins & Slade 1997) identified an appeal to a common set of values as key to the role of narrative in building and reflecting community solidarity. Hunston (1993) notes that the highly implicit evaluations present in professional scientific writing are recognised only because they draw on a shared value system in the scientific community for their interpretation. Thompson & Hunston (2000) generalise from this to suggest that consensus-building is one of the functions of evaluative language. Martin (2000) observes that identifying Appraisal in any given text involves a personal response to the linguistic cues in that text, and acknowledges that other readers may have a different response; this suggests that members of one community may interpret evaluative resources differently from members of another. The obverse of this observation is that evaluative resources play a significant role in constructing common ground and therefore community. Coffin & O'Halloran (2006) make an argument for values to be established intertextually, such that readers of a given newspaper (*The Sun* in this case) gradually build up a shared sense of the value attached to entities that are regularly reported in the newspaper.

In short, one of the identifying features of communities is that they share value systems, and one of the roles of texts that draw on particular value systems is to construe, virtually, a community. We extend this research by adopting a novel methodology for identifying those systems of value and for classifying the instances of evaluative language identified in our corpus. This methodology will be explained below.

Although we adopt a purely social, rather than cognitive, view of language in this paper, we note that our approach is broadly compatible with that of Cognitive Linguistics, and particularly the emerging discipline of Cognitive Sociolinguistics (see, for example, Pütz *et al.* 2014). Kristiansen & Geeraerts (2013: 1) comment that Cognitive Linguistics itself embodies "a contextualised, pragmatic approach to meaning". They add that Cognitive Linguistics is a "usage-based approach to language" and define this in terms of "the dialectic nature of the relation between language use and the language system" (Kristiansen & Geeraerts (2013: 2)). Our argument below that the use of adjectives in the *RMP* community construes a unique system of evaluative meaning is consistent with this. Schönefeld's (2013) is one of a number of studies that demonstrate that co-occurrence of features, in this case specific past participles/adjectives occurring in the '*go un-V-en*' construction (e.g. *went unreported*, *go unnoticed*) can be heavily dependent on register variation. She notes that relevant lexical items occur with differential frequency in each of four broad registers (Conversation, Fiction, News Reporting and Academic Prose), suggesting that the cognitive construct involved is sensitive to social context. Our study involves one register only, but the grounded, bottom-up and discourse specific configuration of meaning that we propose is consistent with this view.

The next section of the paper seeks to problematise taxonomies of evaluative meaning. The subsequent sections explain the method used to derive clusters of adjectives from a corpus compiled from the *RateMyProfessors.com* website, and discuss the resulting categorisation in terms of the 'voices' or personae used by the website contributors. The paper concludes with a discussion of the relevance of this to evaluative taxonomies in general.

3. Taxonomies of evaluative meaning

Approaches to the study of evaluative meaning include those associated with stance (e.g. Biber 2006), evaluation (e.g. Thompson & Hunston 2000), appraisal (e.g. Martin & White 2005) and sentiment (e.g. Liu 2010). A common factor in all these approaches is the establishment of taxonomies of evaluative meaning. The interpretation of individual instances in terms of meaning distinctions is important for both practical and theoretical reasons. Most studies of evaluative language in text make quantitative statements, often comparing corpora in terms of the amount of evaluation of specific kinds found in each (e.g. Bednarek 2008b; Biber 2006; Charles 2006; Fuoli 2012; Hardy & Römer 2013; Hyland 2012). Comparisons of quantity depend on classifications (e.g. 'More use is made of *this* kind of language resource than *that* one in these texts' or 'Corpus/Text A makes more use of *this* kind of resource than does Corpus/Text B').

Attempts to categorise instances of stance, appraisal or evaluation are informed by the need to devise a taxonomy that will handle large numbers of naturally-occurring instances of widely divergent forms. A basic distinction between negative and positive polarity is often used, for example in Sentiment Analysis (e.g. Turney 2002). This presupposes a parameter of value or desirability. Conrad & Biber's (2000) study of stance adverbials divides them, in terms of meaning, into 'epistemic', 'attitudinal' and 'style' domains. Bednarek (2008a) notes that earlier work by Biber & Finegan (1988) uses six categories of stance adverbials, and that Francis (1995) suggests eight categories of evaluative adjectives. Thompson & Hunston (2000) propose four parameters: 'good-bad', 'certainty', 'expectedness' and 'importance'. Bednarek (2008a) herself offers the largest number of parameters — ten — each of which has a number of 'sub-values', ranging from simple polar opposites (e.g. 'comprehensible' and 'incomprehensible' as values of 'comprehensible') to a broader set of possibilities (e.g. 'hearsay', 'mindsay', 'perception', 'general knowledge', 'evidence', 'unspecific' as values of 'evidentiality').

Undoubtedly, the most theorised approach to the categorisation of evaluative language is that developed by Martin, White and others (e.g. Martin & White 2005). Their concept of Judgement, the evaluation of human behaviour, is the category with which we are primarily concerned in this paper. Within that category, a major distinction is made between social esteem and social sanction (Martin & White 2005: 52). This responds to a distinction between the building of social networks through a recognition of shared values (social esteem) and the codification of those values in a legal or moral code (social sanction). Martin & White (2005: 53) comment, wryly, that this distinction might be compared to the distinction between venial and mortal sins in Catholic theology.[1] The authors also align parameters of Judgement with Halliday's model of modalisation (Martin & White 2005: 54, citing Halliday 1994), such that, for example, normality is associated with usuality and propriety with obligation. This alignment is just part of the underpinning of the Appraisal system as an integral aspect of SystemicFunctional Grammar. More generally, the whole concept of Appraisal is based on the bedrock of SFG (e.g. Halliday 1978): that the meaning of a choice from among systemic possibilities depends on what other possibilities are available. Thus the meaning of the selection of an appraisal of social sanction, for example, depends upon the existence of a distinction between social sanction and social esteem.

In general, however, taxonomies of evaluation are open to debate. On the one hand there is, as noted above, disagreement regarding the number of categories required to account for all instances of stance, appraisal or evaluation. On the other hand, there is agreement that, given a set of language resources, a generally satisfactory classification is possible. For example, we can take the pattern '*it is ADJECTIVE to-infinitive*', and identify the most frequent adjectives occurring in the pattern. Using the 400 million word Bank of English corpus, the twenty most frequent adjectives in this pattern are: *hard, difficult, possible, important, easy, impossible, necessary, better, best, good, easier, likely, interesting, safe, essential, nice, great, enough, wrong, fair*. Recurring meanings might be identified as: 'difficult' (*hard, difficult, easy, easier*); 'possible' (*possible, impossible, likely*); 'important' (*important, necessary, essential*); 'right' (*better, best, good, safe, nice, great, wrong, fair*). Each of these groups identifies near-synonyms (*hard* and *difficult; good* and *nice*) and antonyms (*possible* and *impossible; wrong* and *fair*). Some identify words that might be said to be similar in meaning when taken out of an immediate context (e.g. *necessary* and *essential*); in other cases the meaning is similar only when the relevant co-text is considered (e.g. *safe* and *fair* as in *it is safe to say that…* and *it is fair to say that…*).

[1] "We call the most serious and grave sins, mortal sins…. By their very grave nature, a mortal sin cuts our relationship off from God… The second type of sin, venial sin, that of less grave matter, does not cut us off from Christ…" www.saintaquinas.com/mortal_sin.html. Accessed 8th August 2014.

The 'right' group, in particular, suggests that taxonomies are fuzzy, with individual instances of language use blurring the distinction between conceptual categories. The complexities of the relationship between taxonomic category and individual lexical item have been explored in a number of ways. Martin & White (2005: 67–68) approve of double coding in cases where different categories of attitude are inscribed and invoked by a single expression. They give examples such as *he proved a fascinating player*, which inscribes Judgement but invokes Appreciation, and *it was a fascinating innings*, which, conversely, inscribes Appreciation but invokes Judgement. Arguably, both examples also invoke Affect (*I was fascinated by the innings he played*). Bednarek (2006, 2008a) makes the insightful point that single lexical items might express meanings that are bundles of parameters rather than single parameters. The examples she gives include *clear / clearly*, which expresses positive values for both comprehensibility and reliability, and *serious / critical / grave*, which express both reliability and emotivity. Following Lemke (1998) she refers to this phenomenon as 'evaluative conflation'. It is a useful reminder that words do not necessarily map directly on to concepts. In a similar vein, Hunston (2011) identifies what she calls 'multilayering', where a single phrase expresses a number of evaluative meanings. For example, the clause

(1) They are desperate to cling on to power.

can be paraphrased as: 'They wish to retain power' and 'They evaluate power as desirable' (Affect), and as 'I evaluate them as desperate' and 'I evaluate them retaining power as undesirable' (Judgement). Thus, the adjective phrase *be desperate to* might be said to express multi-layered or conflated meanings and is therefore difficult to assign to a single category. Current research by Su (in prep.) suggests that this may be a more common phenomenon than might be expected. For example, of 66 adjectives followed by the preposition *of*, about two-thirds express Affect (e.g. *afraid of, tired of, fond of, ashamed of, desirous of*) and about one-third express Judgement (e.g. *capable of, guilty of, considerate of,*), but a considerable number arguably express both what someone is reported to feel, and how that feeling is judged (e.g.

dismissive of, envious of, protective of, scornful of, tolerant of). Indeed it might be said that it is the norm for Judgement to be invoked whenever Affect is inscribed. This in turn suggests that imposing totally discrete categories is at odds with actual language use.

In short, taxonomies of meaning are necessary both to the practice of analysing evaluative meaning in texts and to the establishment of a theory of evaluative meaning, and yet the expression of evaluative meaning, in English, resists such classifications. Taking adjectives as an example, whereas it is not particularly difficult to identify small groups of adjectives that express similar meanings, it is extremely difficult to categorise unambiguously large numbers of adjectives. While all categorisation might be said to have an element of subjectivity, the extent of the indeterminacy in this area suggests that methodologies that either support or replace a subjective categorisation would be valuable.

Before turning to the methodology to be used in this paper, the point about adjective classification will be illustrated using a corpus drawn from the *RMP* website. (The method of compiling the corpus and identifying the relevant adjectives is described below.) Table 1 lists fifty of the most frequent adjectives describing instructors.[2]

Table 1. Fifty of the most frequent adjectives describing instructors

great	*fair*	*sweet*	not_bad	**not_good**
nice	*knowledgeable*	*fun*	*available*	**terrible**
helpful	*clear*	*excellent*	**not_helpful**	*cute*
good	*interesting*	**tough**	*friendly*	*understanding*
funny	*cool*	*hilarious*	**difficult**	**not_clear**
awesome	*passionate*	**rude**	*interested*	**bad**
willing	**boring**	*enthusiastic*	*hot*	**disorganized**
easy	*smart*	**horrible**	*ok*	*fantastic*
amazing	*wonderful*	*approachable*	*entertaining*	young
hard	*intelligent*	**unclear**	*brilliant*	**unorganized**

[2] As explained in 'Retrieval of adjectives' below, some pre-selection is involved before a set of adjectives is obtained. This explains why we list here 'fifty of the most frequent adjectives' rather than 'the fifty most frequent adjectives'.

The list includes both positive and negative polarity adjectives (positive shown in italics, negative in bold). It is noticeable that most adjectives are very easily assigned to the 'positive' or 'negative' group, with the possible exception of *not_bad* and *young*, where some assumptions are made about likely contexts. The adjectives can also be allocated to groups based on similarity or antonymy in meaning. Our first attempt to do this, using a top-down, researcher-led method of grouping adjectives with similar meaning, yields 15 groups (A-O below).

A. Positive: *great, nice, good, awesome, amazing, cool, wonderful, excellent, fantastic, not_bad, ok*; Negative: *horrible, not_good, terrible, bad*

B. Positive: *helpful, willing;* Negative: *not_helpful*

C. Positive: *funny, fun, hilarious*

D. Positive: *easy;* Negative: *hard, tough, difficult*

E. Positive: *fair*

F. Positive: *knowledgeable, smart, intelligent, brilliant*

G. Positive: *clear;* Negative: *unclear, not_clear*

H. Positive: *interesting, entertaining;* Negative: *boring*

I. Positive: *passionate, enthusiastic, interested*

J. Positive: *sweet, friendly, understanding*

K. Negative: *rude*

L. Positive: *approachable, available*

M. Positive: *hot, cute, young*

N. Negative: *unorganised, disorganized*

Group A is a 'general' group, where the meaning of individual adjectives is non-specific except for the polarity. In a textual context, the meaning might be left vague ('He is wonderful') or specified by a noun ('She is a wonderful teacher') or by a prepositional phrase ('She is wonderful at explaining things'). The other groups denote particular qualities. Our intention in deriving these groups was to employ a minimum amount of interpretation, and this accounts for the large number of groups, but even so it proves impossible to avoid some interpretations of meaning: *sweet* has been taken as similar in meaning to *friendly*; *smart* has been allocated its US sense of 'intelligent' rather than its British sense of 'well-dressed'; *brilliant* has been interpreted in its specific sense of 'intelligent' rather than its general sense of 'excellent'. More contentiously, perhaps, a group has been made of *passionate, enthusiastic* and *interested*, with a quality of emotion as the common factor. Amalgamations of the groups would be possible: for example, group L could be joined with group J, assuming that 'friendliness' and 'approachability' are similar concepts; both could combine with B (*friendly, approachable, helpful*) to give a group relating to personal, as opposed to intellectual, qualities; groups D and G could be combined, assuming that clarity and ease of understanding are similar; groups C and H could be combined as expressing qualities that impact on the evaluator. Such groupings are very far from clear-cut, however. For example, group I (*passionate, enthusiastic, interested*) could be interpreted as an aspect of intelligence (group F), or as an interactive quality (aligned with group C or group J), or as an emotional quality (aligned with group H).

In short, given the list of adjectives in Table 1, it might be agreed that a range of judgements are indicated, including both general (*awesome*) and specific (*rude*) adjectives, and adjectives relating to intellectual (*knowledgeable*) and personal (*friendly*) qualities, as well as those relating to age and appearance (*young, hot, cute*). The actual parameters, and the precise grouping of the adjectives, however, are clearly subjective and might be open to considerable dispute. Given that only one-third of the adjectives have been classified, adding more examples to the list would only make the problem worse. In the next section we consider a bottom-up statistical approach to obtaining categories of linguistic features.

4. Background to the methodology

Evaluative meaning draws heavily on the word class known as adjectives. As Martin & White (2005: 58) succinctly put it: "the canonical grammatical realisation for **attitude** is adjectival".

For this study, then, we focus on adjectives, while recognising that in doing so we will identify only some of the appraisals performed in the corpus.

The methodology breaks with some previous applications of Appraisal theory in that it is quantitative first and qualitative only second. It forms a contrast with work such as Fuoli (2012), wherein Appraisal resources are identified through close analysis of individual texts and then quantified. This is because our first concern is to identify groups of adjectives that consistently co-occur when individual instructors are appraised, rather than to quantify instances of Appraisal based on an *a priori* classification. Our approach is based largely on the Meaning Extraction Method: a quantitative method developed in social psychology to identify recurring meanings or themes in a large number of text samples.

In the first published study using the Meaning Extraction Method (Chung & Pennebaker 2008), a corpus of open-ended self-descriptions written by US college students was analysed. From these texts high frequency personality adjectives were first identified (e.g. *nice, bad, cold, reserved*) and then tabulated to represent their occurrence or absence in individual texts. The purpose of the subsequent processes was to identify how these adjectives tend to co-occur within the self-descriptions. Using a type of factor analysis (Principal Components Analysis), Chung & Pennebaker extracted a seven-factor solution in which the adjectives clustered to form coherent themes. For example, adjective grouping seemed to reflect traits such as sociability (*quiet, shy, reserved, outgoing*) and negative emotions (*sad, hurt, angry, afraid*). The authors argue that when applied to self-descriptions the Meaning Extraction Method was able to reveal psychological dimensions along which authors think about themselves. This approach has subsequently been applied to identify themes in a range of text types including blog posts (Argamon *et al*. 2007), postings on Internet message boards for depression (Ramirez-Esparza *et al*. 2008), emails as part of a patient after-care programme (Wolf *et al*. 2010), and Facebook status updates (Kramer & Chung 2011).

The Meaning Extraction Method has an antecedent in Multi-Dimensional Analysis (e.g. Biber 1988, 1993). Broadly speaking, the aim of Biber's approach is to find statistical co-occurrences between language features in a set of texts, and so distinguish between 'registers'.[3] Each group of co-occurring features (a 'factor') is subsequently interpreted as having a consistent language function (a 'dimension'). For example, Biber (1988: 89) discovers that in texts where simple past tense is frequent, third person pronouns, perfect aspect, public verbs, synthetic negation and present participial clauses are also frequent, whereas present tense verbs and attributive adjectives, among other features, are relatively infrequent. Other language features may show no significant presence or absence and are therefore not included in the cluster of co-occurring features. Having established this list of features that significantly attract or repel each other, Biber identifies it as a 'factor' consisting of the positive and negative attributes (e.g. presence of past tense, absence of present tense). Individual texts, and sets of texts, can be plotted as scoring relatively high or low on this factor. Biber (1988: 108) interprets the factor as indicating the presence or absence of 'narrative action' and labels the dimension 'narrative versus non-narrative concerns'. In other words, a set of features that belong to a particular category (the 'narrative' dimension) are identified, but the method of doing so depends on their mutual co-occurrence rather than on the interpretation of individual texts. Both in the case of Multi-Dimensional Analysis and the Meaning Extraction Method, subjectivity comes into play in the imposition of a label (e.g. 'narrative' or 'sociability') onto the cluster of features identified in the factor, but the categorisation itself is the product of an automated process.

There are a number of reasons why adapting the Meaning Extraction Method to a study of evaluative adjectives might be useful. Primary among these is that using quantitative methods to establish co-occurrence breaks through the cycle of subjectivity that muddies the taxonomic waters in discussing evaluative meaning. This in turn makes it possible to study large amounts of text, taking into account large numbers of adjectives. In contrast to Biber's methodology, in the Meaning Extraction Method features are permitted to occur in more than one factor/dimension. In principle, the appearance of a feature in more than one factor or cluster

[3] It should be noted that the concept of register used by Biber and others (see, for example, Biber & Conrad 2009) is less theorised than that used by Halliday.

demonstrates the non-discrete nature of the categories. It will be argued below that the results are useful in throwing new light on the interaction between adjective classification and community.

In the study reported in this paper, a technique comparable to Chung & Pennebaker's (2008) Meaning Extraction Method is used to cluster adjectives found in the *RateMyProfessors* database to arrive at a classification that is unique to that database. This is described in some detail below, although a fuller account of the methodology is available on request from the authors.

5. Method

5.1 Data selection

RateMyProfessors.com (hereafter *RMP*) is a popular educational review website on which university students can anonymously evaluate their instructors by providing a numerical score (1-5) in three main categories (*helpfulness, clarity, easiness*). From these numerical scores, an 'overall quality' score (*good*, *average*, *poor*) is calculated for each instructor by taking the mean of the scores assigned to that individual for *helpfulness* and *clarity*. Importantly for the present study, students can also optionally write a free-text comment. Founded in 1999, *RMP* now contains about 15 million evaluations of approximately 1.7 million instructors at over 8,000 institutions in the USA, Canada and the UK (RMP 2013). The website was chosen as an appropriate source of texts for this study because it is homogenous in terms of topic (university instructors), and is highly evaluative in nature. In Martin & White's terms, it could be expected to include a large amount of Judgement. Previous research on the numerical scores posted on the site (Sanders *et al*. 2011) demonstrates that they show internal reliability, suggesting that users of the site broadly agree on the positive and negative characteristics of the various individuals appraised and therefore, in the terms discussed above, form a virtual community.

A corpus of the text comments posted on *RMP* was compiled by selecting the top 200 'Best Colleges' as identified in *U.S. News* (2011). From each of these colleges, a random sample of instructors employed there was selected, and the comments relating to those instructors were downloaded. In total 467,904 comments about 50,316 individual instructors were downloaded — this amounted to about 24.6 million words of running text. Automatic annotation for parts of speech using the CLAWS software package (Garside & Smith 1997) enabled the identification of about 1.9 million tokens tagged as adjectives. Because not all these adjectives are relevant to our study, a further stage in retrieving relevant adjectives was implemented.

5.2 Retrieval of adjectives

The retrieval process sought to distinguish those adjectives that index student evaluation of instructors from adjectives performing other roles. In this respect, our application of the Meaning Extraction Method differs considerably from prior studies that have treated the features under analysis (up until now, adjectives and nouns) as a ‘bag of words’ and have disregarded the context in which those features occur. Treating all instances of the same word form as having the same meaning and function is an oversimplification that is likely to detract from the quality of groupings yielded by the statistical analysis. As illustrated by the two examples below, adjectives in the corpus do not necessarily refer to instructors only. In (2), *stupid* describes the instructor, and is identified as a target form in our study; in (3), *stupid* describes a hypothetical student and is excluded from our study.

(2) He is *stupid*, you don't want a professor like this one.

(3) If you get a B, you are seriously *stupid*.

Therefore only adjectives in an attributive or predicative syntactic relationship with a noun or pronoun referring to the instructor, such as in (2), were extracted. These were retrieved by searching for the following abstract string within sentence boundaries:

[proper noun|*he*|*she*] + BE + [up to four intervening words]$^{\text{optional}}$ + ADJECTIVE

This automated process was supplemented by extensive manual checking and annotation of all 328,550 instances of retrieved adjectives, to leave 300,069 tokens indexing student evaluation of their instructor. Spelling mistakes and variations were standardised, and proper adjectives (e.g. *French*, *American*), those adjectives that did not express an evaluation, and other false positives were manually removed.

We also sought to account for the context of adjectives by annotating for the presence of negation. The prefix *not_* was added to the annotation of adjectives that were preceded by adverbs of negation (e.g. *not*, *n't*, *never*, *barely*, *hardly*). The adjectives in (4) were thus labelled *not_old* and *not_annoying*.

(4) She's *not old* and *annoying*.

Although the manual checking was somewhat laborious, it was considered essential to avoid substantial errors entering into the subsequent automatic analysis. These processes resulted in a very large matrix showing how 1,856 adjective types were distributed across 41,633 individual instructors.

5.3 Principal Components Analysis

Our task was then to identify a meaningful structure within this complex dataset. As noted above, following procedures outlined in Chung & Pennebaker (2008) and Wolf *et al*. (2010), we applied the statistical technique called Principal Components Analysis (PCA). PCA is a method for 'simplifying' a multi-dimensional dataset, which, in turn, can help in the identification of a meaningful underlying structure. The general aim of PCA is to "find a small number of linear combinations of the variables so as to capture most of the variation in the dataframe as a whole" (Crawley 2007: 731). Our assumption was that there would be broad agreement between writers, so that groups of adjectives describing an instructor would tend to reflect shared perceptions. This of course is only true where writers have demonstrated that they regard an instructor positively or negatively as evidenced by the numerical ratings given. For this reason the texts commenting on each instructor were classified additionally according to the quality rating accompanying the comment: ‘good’, ‘average’ or ‘poor’. The effect of this is illustrated by (5) and (6), which demonstrate how two students can differ in their evaluation of the same instructor (the actual name of the instructor has been replaced by a pseudonym).

(5) Professor Plum is extremely *clear* and *willing* to go over things with you alone, despite the size of the class. (overall quality = ‘good’)

(6) Professor Plum is an *obnoxious* lecturer and *unresponsive* to student concerns. (overall quality = ‘poor’)

If Professor Plum was treated as a single entity, the resulting group of co-occurring adjectives would include *clear, willing, obnoxious* and *unresponsive*. A group such as this would not represent consensus or establish perceived common ground. To avoid instances such as these, as noted above, instructors were subdivided further according to the ‘overall quality’ rating associated with the comments, so that each instructor could, in theory, have three levels (‘good’, ‘average’, ‘poor’). In the Professor Plum example, Plum is treated as two distinct entities: (1) ‘*PLUM_good_quality*’ (within which the adjectives *clear* and *willing* (5) are tabulated); and (2) ‘*PLUM_poor_quality*’ (within which the adjectives *obnoxious* and *unresponsive* (6) are tabulated).

The data was then tabulated to indicate how many times each adjective is used to describe each instructor+quality ‘entity’ (*PLUM_good; PLUM_poor* etc). However, we are interested in the distribution of adjectives across instructors rather than in the frequency of adjectives and we wish to avoid the skewing effect of different number of comments assigned to each instructor. Therefore, following Chung & Pennebaker (2008), frequencies were replaced by '1' to indicate occurrence or '0' to indicate absence. The outcome of these processes was a very large sparsely populated binary matrix (1856 x 64932) indicating absence or presence of adjectives in comments about each ‘entity’. By treating the many and diverse adjectives as variables, our aim was then to reduce groups of correlated adjectives to a smaller, more manageable number of new variables, or *principal components*.

Prior to carrying out PCA, we reduced the size of input matrix. In the full matrix described above only 0.2% of all cells were populated. As adjectives used to describe very few instructors, and instructors described with very few adjectives, would provide little information for PCA, we excluded low frequency adjectives and instructors with few adjectives; only those adjectives occurring in the texts describing 100 or more instructors were retained, and only those instructors occurring with two or more adjectives. The resulting matrix consisted of 182 columns for adjectives, and 30,000 rows for individual instructors + quality. This matrix formed the initial input for PCA. All analyses were conducted in the R statistical programming environment using the *psych* package (Revelle 2013).

Simple PCA followed by varimax rotation was performed to find clusters of co-occurring adjectives among the 182 frequent adjectives used to describe university instructors. Although we report the results using varimax rotation, other methods of rotation available in the psych package (e.g. quartimax, promax, simplimax) provided results that are virtually identical. Standard tests indicated that the correlation matrix was suitable — the Kaiser-Meyer-Olkin measure (0.82) and Bartlett's test of sphericity (chi-square = 128,083.9, $p < 0.001$). The PCA transformed the 182 original variables (adjectives) into 182 new variables (principal components). However, our aim was to extract only the first few principal components accounting for the most variance in the data (for a fuller explanation of this see Crawley 2007).

We first examined the eigenvalues associated with each principal component. These indicate the amount of total variance explained by each component and increase as variance explained increases. An eigenvalue of greater than 1 indicates that the component accounts for more variance than the original variable and is commonly used as cut to identify components for further investigation (Norman & Streiner 1994). In total 72 components with an eigenvalue of greater than 1 were identified. We also used a scree plot to help decide on the possible numbers of components to retain. The scree plot displays the eigenvalues for each component in descending order. An 'elbow' on the plot indicates a rapid drop in the proportion of variance explained and can help determine how many components to extract. In this case, the scree plot suggested either a solution with two, seven or ten principal components. We therefore repeated the PCA using a varimax rotation, examining two, three, four, five, six, seven, eight, nine and ten component solutions and examined the component loadings in order to find the solution that made the most intuitive sense. Analogous to a correlation coefficient, the loading is a value between 1 and -1 that indicates the strength of relationship between the original variables (i.e. the adjectives) and the new principal component. The seven-component solution was chosen as strongly loaded adjectives formed intuitive clusters. This solution is presented in full in Appendix 1.

In the final step we simplified this solution. By examining how the individual adjectives load onto each of the seven components, we identified those adjectives with component loadings of less than 0.2 and more than -0.2. These made little or no contribution to the overall variation in the dataset and were removed from the input matrix. In total we removed 101 adjectives resulting in a simplified input matrix (81 x 30,000). Again, PCA followed by varimax rotation was performed with standard tests indicated that the correlation matrix was suitable — the Kaiser-Meyer-Olkin measure (0.84) and Bartlett's test of sphericity (chi-square = 85,736.1, $p < 0.001$). Twenty-three components had eigenvalues of greater than one. Based on the scree plot the first seven components were extracted — this solution is presented in full in Appendix 2.

The final seven-component solution accounted for just under 16% of the total variance — the variance explained by the individual components is shown in Table 2.

Table 2. Variance explained by the seven-component solution

	Variance explained by component	Cumulative variance explained
PC1	4.80%	4.80%
PC2	2.67%	7.47%
PC3	1.85%	9.32%
PC4	1.73%	11.05%
PC5	1.64%	12.69%
PC6	1.60%	14.29%

PC7	1.58%	15.87%

The groupings of adjectives in this seven-component solution represented a balance between interpretability and variance explained. Of course, 16% is relatively low, and applied to other types of data PCA would often account for a larger proportion of variance. However, the figure is not unusual for the analysis of natural language. The proportions are in line with comparable studies; e.g. between 9% and 14% in (Chung & Pennebaker 2008) and 12% in Wolf *et al*. (2010). Factors that contribute to the low proportion of variance include the uneven distribution of the adjectives over individual instructors (only 6.1% of cells in the final input matrix were populated), and the productive nature of language (comparable characteristics can be expressed in numerous different ways).

For each of the seven principal components, the most strongly associated adjectives were identified by reference to the component loading. Of the 81 adjectives, 78 occur with component loadings of 0.2 and higher, and 17 of these adjectives load on more than one principal component. The adjective groupings that emerge from this process are presented and discussed in the following sections.

6. Results

Table 3 shows the adjectives in each group with a loading of 0.2 or above, with figures indicating the strength of co-association,[4] with figures indicating the strength of co-association.

[Please insert Table 3 here]

[4] As we are concerned with co-association, this table excludes three adjective types with **negative** loadings (i.e. less than -0.2) — *helpful* and *great* are excluded from group 3, and *nice* from group 4 and 5.

Effectively, each group consists of adjectives that writers use in combination with each other. In other words, cohorts of writers who describe a particular instructor (positively) as *helpful* (group 1) are likely also to describe them as *willing*, *great* and *sweet*, thus associating 'sweetness' with 'helpfulness'. Cohorts of writers who describe the same or another instructor (negatively) as *rude* (group 4) are likely also to describe them as *condescending* or *arrogant*, thus associating 'arrogance' with 'rudeness'. These co-occurrences are in some ways at odds with the intuitive grouping offered above. According to that classification, *helpful, willing, sweet* and *great* come in three separate groups (A, B and J). The co-occurrences shown in table 3 also include adjectives which are less frequent but which show strength of co-occurrence. For example, *condescending* and *arrogant* are not frequent enough to appear in the top 50 list shown in table 1, but strongly associate with *rude*.

Thus, the grouping suggested by Principle Components Analysis offers an alternative interpretation of the adjectives in the corpus. These groups will now be used as the basis for a qualitative interpretation below.

7. Discussion: A community-based taxonomy

The classification in Table 3 is somewhat more fluid than the A-N groups suggested above. For example, *clear* is found in group 1, associated with *helpful*, and also in group 6, associated with *tough*. The adjectives *smart* and *intelligent* co-occur with adjectives such as *brilliant* (group 5) as well as with *tough* and *difficult* (group 6). The adjective *hard* occurs in group 3, with *horrible* and in group 6, with *difficult* but also with *intelligent*. Most of the general adjectives (group A above) occur in more than one group (*great, nice, good, awesome, amazing, cool, wonderful)*. Groups 3 and 4 both contain the adjectives *not_helpful, awful* and *useless*. These varying co-occurrences suggest that the meaning of the adjectives may be nuanced in particular ways. For example, *great* might indicate a caring nature (1), or good humour (2). The quality of intelligence (*intelligent, smart*) is associated with both knowledge (5) and challenge (6). That of clarity (*clear*) is nuanced towards either a personal quality (helpfulness, group 1) or a professional strategy (challenge, group 6).

For the most part, the groups are either positive in polarity (groups 1, 2, 5, 7) or negative (groups 3, 4). This is not surprising, as groups of texts associated with negative and positive numerical scoring have been separated, as described above. However, groups 2 and 6 include both positive adjectives (*entertaining, fair*) and negative ones (*sarcastic, difficult*). These are not antonyms of one another, as *possible* and *impossible* are, but rather include positive and negative perspectives on a single quality (*entertaining* and *sarcastic*), or indicate qualities that might be said to be complementary (*difficult* but also *fair*). This might be explained pragmatically rather than semantically. Writers describe their instructors (positively) as 'tough but fair' but not as 'funny but boring' or 'clear but unclear'. In other words, the groups of co-occurrences are socially-driven rather than semantically-driven. There is no reason to suppose that similar groups would be found if other communities were recording judgements; rather, they are instantial and specific to this corpus.

To aid further discussion of the seven groups, mnemonics will be used to indicate the main focus of each group, as follows:

1. 'HELPFULNESS' *helpful, willing, sweet, caring, available, understanding, approachable…*
2. 'FUNNINESS' *funny, hilarious, entertaining, fun, sarcastic, crazy...*
3. 'INCOMPETENCE' *horrible, unclear, terrible, confusing, unorganised, boring,…*
4. 'RUDENESS' *rude, condescending, arrogant, unhelpful, mean…*
5. 'INTELLIGENCE' *passionate, brilliant, intelligent, knowledgeable, , inspiring, engaging, interesting, smart …*
6. 'DIFFICULTY' *tough, hard, difficult, fair, intimidating, clear…*
7. 'HOTNESS' *hot, gorgeous, young, beautiful, attractive…*

Of the seven groups, one relates to appearance ('HOTNESS'), three relate to generally positive character and behaviour attributes ('HELPFULNESS', 'FUNNINESS', 'INTELLIGENCE'), one is less easy to define in terms of polarity ('DIFFICULTY') and two relate to unambiguously negative attributes ('INCOMPETENCE' and 'RUDENESS'). In terms of Appraisal, the HOTNESS group may be categorised as Appreciation while the others relate to Judgement. The three 'positive' and one 'mixed' Judgement groups together reflect two specific parameters: intellectual ability ('how clever'?) represented by just one group, INTELLIGENCE, and interpersonal ability ('how skilled as a teacher' or 'how good as a person'?) represented by three: HELPFULNESS, FUNNINESS and DIFFICULTY. The negative groups divide interestingly between 'social sanction' (RUDENESS) and 'social esteem' (INCOMPETENCE), although, as noted above, there is some overlap between them. The mnemonic 'RUDENESS' does not itself indicate 'mortal sin', to quote Martin & White (2005), but the group includes adjectives such as *disrespectful, unapproachable* and, most of all, *unprofessional,* which suggest a level of poor behaviour that might incur sanctions from an employer. For the writers of these texts, then, 'rudeness' and its associated qualities suggest a harsher judgement than the general population might assign to it. The INCOMPETENCE group, on the other hand, includes adjectives such as *boring, confusing* and *disorganised*; arguably these suggest personal rather than professional failings — venial rather than mortal sins. Somewhat surprisingly, then, this community associates 'rudeness' with 'unprofessionalism', whereas 'incompetence' appears to be less severely sanctioned.

The nuancing of qualities in this corpus, through the co-occurrence of adjectives, is observable in the 'positive' groups as well. Specifically, the qualities mentioned are not simply inherent in the instructor being evaluated but are inflected towards the 'end-user' (the student). For example, the INTELLIGENCE group includes *passionate*, *inspiring, engaging, interesting* as well as adjectives more clearly related to intellectual ability. In this corpus, then, 'intelligence' is not only about knowledge or intellectual capacity, but also about interaction with the learner, or the novice intellectual. A focus on the student as consumer or audience may be observed also in the FUNNINESS group. Humour is presented as an interactional quality, with *entertaining* and *fun* (how the evaluated person affects others) appearing as well as *hilarious* and *energetic*.

Continuing this argument and relating it to Coffin's (2002) notion of evaluative voices requires us to move away from a strict adherence to the groups identified, and to observe alignments within but also between groups of adjectives. We argue that the community of writers in this corpus sanctions three voices, from which the individual writer chooses: the evaluator as novice intellectual (valuing intelligence and difficulty); the evaluator as consumer (valuing being entertained, interested and helped); and the evaluator as subordinate (valuing fairness and approachability). These are now elaborated upon in turn.

The overlap between the INTELLIGENCE and the DIFFICULTY groups suggests that students are willing also to see themselves as novice intellectuals, interpreting challenge and difficulty positively, as a sign of intellectual ability in the instructor rather than as unhelpfulness towards the student. An interpretation of the INTELLIGENCE and FUNNINESS groups suggests that students writing the texts in the corpus perceive themselves as 'consumers' in the instructor-student relationship, and therefore holding a position of power. Alternatively, students can present themselves as taking a subordinate power position. The DIFFICULTY group, for example, includes *intimidating*. The HELPFUL group suggests an inflection both towards 'consumer' and towards 'subordinate': the adjectives *helpful, caring, kind* and *friendly* suggest the first interpretation (the instructor is in a customer-service role), while *available, patient* and *approachable* suggest the second (the instructor is in a powerful position but allows access nonetheless).

Turning to the negative polarity groups (RUDENESS and INCOMPETENCE), and remembering the 'social sanction' / 'social esteem' distinction discussed above, it may be argued that the 'consumer' voice strongly predominates. This is suggested by the striking association of interpersonal transgressions such as rudeness and arrogance with unprofessionalism and therefore with social sanction rather than esteem. The focus on unhelpfulness and lack of clarity in the INCOMPETENCE group is another indicator of the consumer voice.

It is apparent that the metalanguage associated with Appraisal is extremely useful in accounting for the adjective groups arrived at with this methodology. We are, however, faced with two possible interpretations of our data. One interpretation is that there is a baseline set of categories for Appraisal, from which any texts will select. This would be Coffin's interpretation, as far as we can tell. Unsurprisingly, and as far as we can tell by looking at adjectives alone, the texts in this corpus select Judgement rather than Appreciation, for the most part, and prioritise Capacity. We have suggested that Capacity might be further divided into professional and personal qualities.

The alternative interpretation, however, is that a more meaningful and robust classification of adjectives has been obtained from calculating co-occurrence than from an attempt to obtain discrete categories without that information. If this is the case, the grouping of adjectives in this corpus holds true for this corpus only, or for this community (contributors to *RMP* comment texts). It might be argued that the community selects from a range of general possibilities only in the broadest sense, and instead creates its own range of possibilities. To some extent, of course, the instantial groups created by this corpus are aligned with distinctions proposed elsewhere. The distinction between the HOTNESS group and the others, for example, aligns with the Appreciation / Judgement distinction. However, more insight has been gained by considering the value system shared by the writers of the texts in this corpus as something that is self-contained than by relating it to more general systems.

One consequence of this conclusion is that categories of meaning are fluid and context-dependent. We have expressed this concept of fluidity and dependency in terms of communities: virtual groups who collectively construe a classification of meaning that is valid only for that community. This appears to us to be in line with a social and functional view of language that defines meaning in terms of taxonomies of choice. We have not raised the issue of the individuals who comprise that community or indeed the meaning of the individual adjectives that appear in our lists. A possible inference from this study, however, might be that the meaning of individual adjectives, as well as semantic categories, is also fluid and context-dependent. This would be in line with Teubert's (2010) view that word meaning is to be found only within discourse, but also, perhaps strangely, with the view within Cognitive Linguistics that for individuals the mental representation of meaning is not fixed. Dunbar (2012), citing

Tuggy (1993), for example, argues that whether or not a given lexeme (such as *paint*) is polysemous for an individual (i.e. whether *paint a portrait, paint a wall, paint nails* are 'the same' or 'different') will vary according to context. Liu (2013: 96) suggests that "there is a certain degree of fluidity in … the internal semantic structure of a synonymous-noun set". Whilst we do not speculate on the mental lexicons of the individuals who contribute to *RateMyProfessors.com*, there is some support in this paper for the view that adjectives used to evaluate individuals have very flexible meanings.

8. Conclusion

This paper has presented an innovative methodology for grouping adjectives expressing evaluation of instructors in the website *RateMyProfessors.com*. Adapting Chung & Pennebaker's (2008) Meaning Extraction Method, Principle Components Analysis has been used to establish levels of co-occurrence between adjectives in this corpus, and to derive groups based on that co-occurrence. This methodology has been used as an alternative to allocating adjectives to classes within existing taxonomies, with the aim of establishing groups that are meaningful within the community of writers whose texts are included in the corpus. Seven groups of adjectives have been identified. They have been allotted mnemonic labels and six of them (excluding APPEARANCE) have been discussed in some detail.

Although the groups are recognisably coherent, they are not semantically homogenous. There appears to be contradictory information: some adjectives occur in more than one group, and some groups contain adjectives that are dissimilar to one another. This suggests that the groups are pragmatic rather than semantic in nature, and that evaluative meaning is socially-inflected rather than abstract. The community of *RMP* writers is characterised by particular attitudes, not only to the individuals evaluated but to teacher qualities in general.

We have argued that the seven groups represent evaluative configurations (following Coffin 2002) that we have labelled the 'novice intellectual', 'consumer', and 'subordinate' voices. It is noticeable that the negative groups (RUDENESS and INCOMPETENCE) prioritise the 'consumer' voice, suggesting that this is the role students adopt when negatively evaluating their instructors. The positive groups are more mixed in the voices they represent. The voices represent the personae of the writers of the free-text comments on *RMP* in terms of their relationship with the instructors being evaluated, just as the voices identified by Coffin represent the personae of the student historians in relation to the events they are recounting.

There are applications of our work to discussions about the relevance and reliability of student evaluations of their instructors, whether these are gathered officially by employing institutions, or unofficially, as in *RMP*. The identification of various voices used by the writers, particularly in their negative evaluations, might inform that discussion. To a large extent our work supports the view that evaluators respond to personal characteristics as much or more than to professional expertise. It also suggests that writers, in adopting evaluative configurations of the relationship kind, are influenced by their own performance as scholars as well as that of their instructors.

The study reported in this paper raises the question as to how far the method might be repeated on other datasets (corpora). It might be noted that the *RMP* website has a number of features that have greatly assisted the quantitative research. The comment texts on the site have a single, fairly narrow, function, that of recording evaluations of professors. A very large number of such texts are available for investigation, and that investigation is helped by the fact that the writers have also given numerical scores for the professors they evaluate, allowing positive and negative views of a single individual to be grouped separately. These features made *RMP* a particularly good dataset to select for the study, but they are unlikely to be present in many other opinion websites.

Notwithstanding these limitations, the paper has offered a quantitative methodology (based on Chung & Pennebaker 2008) for processing very large numbers of texts so that meaningful clusters of adjectives are identified for further qualitative discussion. This does not mean that the methodology was independent of human intervention. An important variation to the Meaning Extraction Method was the way that context was taken into account, in extracting only adjectives in a particular syntactic frame and annotating for negation. This involved extensive manual checking of thousands of instances of adjectives. Further intervention was involved in selecting the number of groups from a small set of possibilities, and of course in providing mnemonic labels for the groups.

There are, then, limits to the extent to which the method could be said to be automatic. It does allow for processing of a much larger amount of data than is possible with more manual methods. We suggest that studies of this kind might complement the more fine-grained analysis of single texts. We have argued that adjective classifications might be community-specific; further research on different corpora will strengthen or challenge that claim.

Submitted 16 March 2014.
Revised version received 24 August 2014.
Revised version accepted 14 October 2014.

References

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker & Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday* 12. Available at: <http://uncommonculture.org/ojs/index.php/fm/article/view/2003/ 1878>. Date accessed: 12 August 2014.

Bednarek, Monika. 2006. *Evaluation in media discourse*. London: Continuum.

Bednarek, Monika. 2008a. 'An increasingly familiar tragedy': Evaluative collocation and conflation. *Functions of Language* 15. 7–34.

Bednarek, Monika. 2008b. *Emotion talk across corpora*. London: Palgrave.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: CUP.

Biber, Douglas. 1993. The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities* 26. 331–345.

Biber, Douglas. 2006. Stance in spoken and written university registers. *Journal of English for Academic Purposes* 5. 97–116.

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge: CUP.

Biber, Douglas & Edward Finegan. 1988. Adverbial stance types in English. *Discourse Processes* 11. 1–34.

Brown, Michael J., Michelle Baillie & Shawndel Fraser. 2009. Rating RateMyProfessors.com: A comparison of online and official student evaluations of teaching. *College Teaching* 57(2). 89–92.

Charles, Maggie. 2006. The construction of stance in reporting clauses: A cross-disciplinary study of theses. *Applied Linguistics* 27. 492–518.

Charles, Maggie. 2007. Argument or evidence? Disciplinary variation in the use of the noun-that pattern in stance construction. *English for Specific Purposes* 26. 203–218.

Chen, Yining & Leon B. Hoshower. 2003. Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment and Evaluation in Higher Education* 28. 71–88.

Chung, Cindy K. & James W. Pennebaker. 2008. Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality* 42(1). 96–132.

Coffin, Caroline. 2002. The voices of history: Theorising the interpersonal semantics of historical discourses. *Text* 22(4). 503–528.

Coffin, Caroline & Kieran O'Halloran. 2006. The role of appraisal and corpora in detecting covert evaluation. *Functions of Language* 13. 77–110.

Conrad, Susan & Douglas Biber. 2000. Adverbial marking of stance in speech and writing. In Susan Hunston & Geoff Thompson (eds.), 56–73.

Crawley, Michael J. 2007. *The R book*. New Jersey, NJ: Wiley.

Davison, Elizabeth & Jammie Price. 2009. How do we rate? An evaluation of online student evaluations. *Assessment & Evaluation in Higher Education* 34(1). 51–65.

Drasovean, Anda & Caroline Tagg. Submitted. Evaluation and its community-building role.

Dunbar, George. 2012. Adaptive Resonance Theory as a model of polysemy and vagueness in the cognitive lexicon. *Cognitive Linguistics* 23(3). 507–537.

Eggins, Suzanne & Slade, Diane. 1997. *Analysing casual conversation*. London: Cassell.

Felton, James, John Mitchell & Michael Stinson. 2004. Web-based student evaluations of professors : The relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education* 29(1). 91–108.

Felton, James, Peter T. Koper, John Mitchell & Michael Stinson. 2008. Attractiveness, easiness and other issues: Student evaluations of professors on Ratemyprofessors.com. *Assessment & Evaluation in Higher Education* 33(1). 45–61.

Fuoli, Matteo. 2012. Assessing social responsibility: A quantitative analysis of Appraisal in BP's and IKEA's social reports. *Discourse and Communication* 6(1). 55–81.

Francis, Gill. 1995. Corpus-driven grammar and its relevance to the learning of English in a cross-cultural situation. Unpublished manuscript.

Garside, Roger & Nicholas Smith. 1997. A hybrid-grammatical tagger: CLAWS 4. In Roger Garside, Geoffrey Leech & Anthony McEnery (eds.), *Corpus annotation: Linguistic information from computer text corpora*, 102–121. Harlow: Longman.

Gee, James Paul. 2005. Semiotic social spaces and affinity spaces: from *The Age of Mythology* to today's schools. In David Barton & Karin Tusting (eds.), *Beyond communities of practice: Language power and social context*, 214–232. Cambridge: CUP.

Halliday, M.A.K. 1978. *Language as social semiotic: The social interpretation of language and meaning*. London: Arnold.

Halliday, M.A.K. 1994. *An introduction to functional grammar*, 2nd edition. London: Arnold.

Hardy, Jack A. & Ute Römer. 2013. Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-level Student papers (MICUSP). *Corpora* 8. 183–207.

Hiltunen, Turo. 2010. *Grammar and disciplinary culture: A corpus-based study*. Helsinki: University of Helsinki.

Hunston, Susan. 1993. Evaluation and ideology in scientific writing. In M. Ghadessy (ed.), *Register analysis: Theory and practice*, 57–74. London: Pinter.

Hunston, Susan. 2011. *Corpus approaches to evaluation: Phraseology and evaluative meaning*. London: Routledge.

Hunston, Susan & Geoff Thompson (eds.). 2000. *Evaluation in text: Authorial stance and the construction of discourse*. Oxford: OUP.

Hyland, Ken. 2005. *Metadiscourse*. London: Continuum.

Hyland, Ken. 2009. *Academic discourse*. London: Continuum.

Hyland, Ken. 2012. *Disciplinary identities: Individuality and community in academic discourse*. Cambridge: CUP.

Iedema, Rick, Susan Feez & Peter R. R. White. 1994. *Media literacy*. Sydney: Disadvantaged Schools Program, NSW Department of School Education.

Knight, Naomi. 2008. 'Still cool…and american too!': An SFL analysis of deferred bonds in internet messaging humour. In Nina Nørgaard (ed.), *Systemic Functional Linguistics in use, Odense Working Papers in Language and Communication*, 481–502. Retrieved 3rd January 2014 from http://static.sdu.dk/mediafiles/Files/Om_SDU/Institutter/ISK/Forskningspublikationer/OWPLC/Nr29/Naomi_Knight.pdf.

Kramer, Adam D.I. & Cindy K. Chung. 2011. Dimensions of self-expression in Facebook status updates. *Proceedings of the 2008 International Conference on Weblogs and Social Media (ICWSM2011), Palo Alto, CA, July 2011*. 169–176.

Kristiansen, Gitte & Dirk Geeraerts. 2013. Contexts and usage in Cognitive Sociolinguistics. *Journal of Pragmatics* 52. 1–4.

Lemke, Jay. 1998. Resources for attitudinal meaning: Evaluative orientations in text semantics. *Functions of Language* 5(1). 33–56.

Liu Bing. 2010. Sentiment analysis and subjectivity. In Nitin Indurkhya & Fred J. Damerau (eds.) *Handbook of Natural Language Processing*. 2nd edn, 627–665. Boca Raton, FL: Chapman & Hall / CRC.

Liu Dilin. 2013. Salience and construal in the use of synonymy: A study of two sets of near-synonymous nouns. *Cognitive Linguistics* 24(1). 67–113.

Marsh, Herbert W. 2007. Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In Raymond P. Perry & John C. Smart (eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective*, 319–383. Springer.

Martin, James R. 2000. Beyond exchange: Appraisal systems in English. In Susan Hunston & Geoff Thompson (eds.), 142–175.

Martin, James R. & Peter R.R. White. 2005. *The language of evaluation: Appraisal in English*. London: Palgrave.

Norman, Geoffrey R. & David L. Streiner. 1994. *Biostatistics: The bare essentials*. St Louis: Mosby.

Polanyi, Livia. 1985. *Telling the American story: From the structure of linguistic texts to the grammar of a culture*. Norwood, NJ: Ablex.

Pütz, Martin, Justyna A. Robinson & Monika Reif (eds.). 2014. *Cognitive sociolinguistics: Social and cultural variation in cognition and language use*. Amsterdam: Benjamins.

Ramirez-Esparza, Nairan, Cindy K. Chung, Ewa Kacewicz, & James W. Pennebaker. 2008. The psychology of word use in depression forums in English and in Spanish. Testing two text analytic approaches. *Proceedings of the 2008 International Conference on Weblogs and Social Media (ICWSM2008), Menlo Park, CA, April 2008*. 102–108.

RateMyProfessors.com. 2013. About us. Retrieved August 6, 2013, from http://www.ratemyprofessors.com/About.jsp

Revelle, William. 2013. *Psych*. Procedures for personality and psychological research, Northwestern University, Evanston, Illinois, USA, http://CRAN.R-project.org/package=psych Version = 1.3.2.

Sanders, Shane, Bhavneet Walia, Joel Potter & Kenneth W. Linna. 2011. Do more online instructional ratings lead to better prediction of instructor quality? *Practical Assessment, Research & Evaluation*. 16(2). 1–6.

Schiffrin, Deborah. 1984. How a story says what it means and does. *Text* 4. 313–346.

Schönefeld, Doris. 2013. 'It is…quite common for theoretical predictions to go untested (BNC_CMH)'. A register-specific analysis of the English *go un-V-ed* construction. *Journal of Pragmatics* 52: 17–33.

Shevlin, Mark, Philip Banyard, Mark Davies & Mark Griffiths. 2000. The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment and Evaluation in Higher Education* 25. 397–405.

Stenglin, Maree K. 2004. *Packaging curiosities: Towards a grammar of three-dimensional space.* Sydney: University of Sydney PhD thesis.

Su Hang. In preparation. Adjectives of judgement in biographical discourse: A corpus study. Birmingham: University of Birmingham PhD thesis.

Teubert, Wolfgang. 2010. *Meaning, discourse and society*. Cambridge: CUP.

Thompson, Geoff & Susan Hunston. 2000. Evaluation: An introduction. In Susan Hunston & Geoff Thompson (eds.), 1–27.

Tuggy, David. 1993. Ambiguity, polysemy and vagueness. *Cognitive Linguistics* 4. 273–290.

Turney, Peter D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, July 2002*. 417–424.

Wolf, Markus, Cindy K. Chung & Hans Kordy. 2010. Inpatient treatment to online aftercare: e-mailing themes as a function of therapeutic outcomes. *Psychotherapy research: Journal of the Society for Psychotherapy Research* 20(1). 71–85.

Zappavigna, Michelle. 2011. Ambient affiliation: A linguistic perspective on Twitter. *New Media & Society* 13(5). 788–806.

Appendix 1

Seven-component solutions for varimax-rotated principal-components analysis of selected 182 adjectives

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
helpful	0.44		-0.16	-0.19		0.14	
willing	0.41			-0.10		0.14	
sweet	0.35		-0.11	0.15			
great	0.34	0.25	-0.13	-0.23		0.11	0.11
wonderful	0.33				0.19		
clear	0.31	0.11		-0.12		0.29	
nice	0.31		-0.21	0.20	-0.17	0.15	0.11
caring	0.30						
amazing	0.29	0.21			0.29		
easy	0.28	0.20			-0.13		0.14
understanding	0.26						
kind	0.25						
patient	0.25						
available	0.25					0.11	
excellent	0.24				0.18	0.19	
approachable	0.21					0.12	
friendly	0.21						
concerned	0.19						
sincere	0.17				0.14		
organized	0.16					0.12	
happy	0.15						

genuine	0.14						
adorable	0.14	0.11					0.12
flexible	0.14						
lenient	0.13						
eager	0.13						
quick	0.12						
generous	0.12						
compassionate	0.12						
considerate	0.11						
personable	0.11	0.10					
respectful	0.11						
perfect	0.10						
accommodating							
pleasant							
effective							
responsive							
funny		0.55					0.12
hilarious		0.49					
entertaining		0.49			0.11		
interesting		0.39			0.21		
awesome	0.31	0.34		-0.15			0.14
fun	0.15	0.33					0.11
cool		0.27					0.23
crazy		0.23		0.14			
energetic	0.12	0.22			0.12	-0.13	
enthusiastic	0.15	0.21			0.19		
quirky		0.20					
sarcastic		0.20	0.16			0.14	
animated		0.19					
witty		0.18			0.12	0.10	
humorous		0.17					
eccentric		0.15					
weird		0.14		0.11			
goofy		0.13					
excited		0.13					
amusing		0.13					
strange		0.12					
enjoyable		0.12					
laid-back		0.11					
charismatic		0.10					
likable							
liberal							
not_boring							
informative							
relaxed							
different							

rude		0.60	0.10		
condescending		0.40			
arrogant		0.40			
unhelpful		0.36	0.14		
mean		0.34			
not_helpful		0.31	0.27		
disrespectful		0.29			
not_nice		0.27			
awful		0.26	0.25		
unwilling		0.23			
pompous		0.22			
useless		0.22	0.19		
unprofessional		0.21			
unapproachable		0.21			
unfair		0.17	0.10		
moody		0.16			
not_willing		0.16			
cocky		0.12			
biased		0.12			
harsh		0.11			0.11
annoying		0.11	0.11		
wrong		0.10	0.10		
anal					
critical					
opinionated					
horrible		0.32	0.38		
unclear		0.16	0.36		
terrible		0.22	0.35		
boring			0.34		0.13
unorganized			0.33		-0.11
confusing			0.33		
not_good		0.11	0.30		
bad		0.12	0.29		
not_clear			0.25		
disorganized			0.24		
old	0.11		0.21		
late			0.18		
ridiculous		0.12	0.17		
poor			0.15		
impossible		0.12	0.15		
vague			0.13		
scatterbrained			0.12		
random			0.12		
slow			0.10		
awkward					
obsessed					

lazy							
not_interesting							
busy							
brilliant					0.39		
intelligent					0.38	0.17	
passionate	0.11	0.15			0.37		
knowledgeable					0.33	0.15	
inspiring					0.28		
insightful					0.27		
smart		0.11		0.11	0.25	0.17	0.12
incredible	0.11				0.25		
engaging		0.15			0.23		
fantastic	0.18	0.11			0.19		
interested	0.16	0.11			0.17		
challenging					0.17	0.16	
fabulous					0.15		
articulate					0.15		
outstanding					0.15		
dedicated					0.14		
talented					0.13		
encouraging	0.12				0.13	-0.10	
supportive	0.11				0.13		
professional					0.12		
open					0.12		
phenomenal					0.11		
honest					0.10		
terrific							
creative							
tough					0.19	0.36	
not_bad				0.11		0.35	
hard				0.19		0.32	
good	0.22	0.20				0.30	0.10
fair	0.19				0.15	0.28	
not_easy						0.22	
difficult				0.15	0.11	0.22	
intimidating					0.12	0.21	
ok				0.16		0.19	
strict						0.19	
concise	0.10					0.18	
straightforward						0.17	
alright						0.15	
decent						0.15	
thorough	0.13					0.14	
not_hard						0.13	
demanding						0.13	
positive	0.11				0.10	-0.13	

specific						0.12	
picky						0.12	
accessible						0.11	
fine						0.11	
dry						0.11	
reasonable							
serious							
not_great							
not_difficult							
hot							0.54
gorgeous							0.46
young							0.41
attractive							0.35
beautiful					0.16		0.35
cute		0.10					0.33
sexy							0.29
pretty							0.22
bright					0.13		0.13
easy-going							0.11
mad							
chill							
new							
understandable							

Appendix 2

Seven-component solutions for varimax-rotated principal-components analysis of selected 81 adjectives

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
helpful	**0.49**		-0.20	-0.13			
willing	**0.46**		-0.11			0.10	
great	**0.37**	0.25	-0.24		0.10		
sweet	**0.37**		0.14			-0.14	
wonderful	**0.36**				0.23		
nice	**0.35**		0.18	-0.21	-0.21	0.11	
clear	**0.34**	0.13	-0.11			0.20	
caring	**0.31**						
easy	**0.31**	0.23			-0.12		
kind	**0.27**						
available	**0.27**						
understanding	**0.27**						
excellent	**0.27**				0.17	0.17	
patient	**0.26**						

approachable	**0.23**						
friendly	**0.23**						
concerned	0.19						
funny		**0.60**					
hilarious		**0.53**			0.11		
entertaining		**0.52**			0.14		
interesting		**0.38**			0.25		
awesome	0.32	**0.36**	-0.14		0.10		0.10
fun	0.13	**0.35**				-0.11	
cool		**0.32**					0.20
sarcastic		**0.24**		0.16		0.15	
crazy		**0.23**	0.15				
energetic	0.11	**0.22**			0.16	-0.15	
quirky		0.18					
horrible			**0.44**	0.27			
unclear			**0.42**				
terrible			**0.39**	0.19			
confusing			**0.37**				
unorganized			**0.36**			-0.15	
not_good			**0.33**				
not_helpful			**0.32**	0.28			
bad			**0.31**			0.13	
boring			**0.31**		-0.11	0.14	
awful			**0.31**	0.23			
not_clear			**0.28**				
disorganized			**0.26**				
old		0.13	0.19			0.12	
rude			0.14	**0.61**			
condescending				**0.47**			
arrogant				**0.45**			
unhelpful			0.18	**0.36**			
mean			0.11	**0.33**			
disrespectful				**0.30**			
pompous				**0.30**			
not_nice				**0.28**			
unwilling				**0.24**			
useless			0.21	**0.21**			
unapproachable				**0.21**			
unprofessional				**0.20**			
passionate	0.12				**0.42**		
brilliant					**0.42**	0.16	
intelligent					**0.39**	0.24	
knowledgeable					**0.35**	0.19	
amazing	0.31	0.20			**0.33**		
incredible	0.12				**0.32**		
inspiring					**0.31**		

insightful					**0.29**		
engaging		0.13			**0.26**		
enthusiastic	0.15	0.18			**0.24**		
fantastic	0.18				**0.22**		
tough					0.11	**0.43**	
hard			0.20			**0.42**	
not_bad					-0.14	**0.39**	
fair	0.21		-0.11		0.12	**0.32**	
good	0.26	0.21				**0.29**	
difficult			0.15			**0.28**	
intimidating						**0.27**	
smart					0.24	**0.25**	0.13
not_easy						**0.25**	
hot							**0.57**
gorgeous							**0.49**
young							**0.42**
beautiful					0.14		**0.40**
attractive							**0.38**
cute		0.13					**0.33**
sexy							**0.31**
pretty							**0.23**

Authors' addresses

Neil Millar
Graduate School of Systems and
Information Engineering
University of Tsukuba
Tsukuba
Japan
(corresponding author)

millar.neil.gm@u.tsukuba.ac.jp

Susan Hunston
Department of English Language
and Applied Linguistics
University of Birmingham
Birmingham B15 2TT
United Kingdom

s.e.hunston@bham.ac.uk

Table 3:

Group 1	
helpful	0.49
willing	0.46
great	0.37
sweet	0.37
wonderful	0.36
nice	0.35
clear	0.34
awesome	0.32
caring	0.31
amazing	0.31
easy	0.31
kind	0.27
available	0.27
understanding	0.27
excellent	0.27
patient	0.26
good	0.26
approachable	0.23
friendly	0.23
fair	0.21

Group 2	
funny	0.60
hilarious	0.53
entertaining	0.52
interesting	0.38
awesome	0.36
fun	0.35
cool	0.32
great	0.25
sarcastic	0.24
crazy	0.23
easy	0.23
energetic	0.22
good	0.21

Group 3	
horrible	0.44
unclear	0.42
terrible	0.39
confusing	0.37
unorganized	0.36
not_good	0.33
not_helpful	0.32
bad	0.31
boring	0.31
awful	0.31
not_clear	0.28
disorganized	0.25
useless	0.21
hard	0.20

Group 4	
rude	0.61
condescending	0.47
arrogant	0.45
unhelpful	0.36
mean	0.33
disrespectful	0.30
pompous	0.30
not_helpful	0.28
not_nice	0.28
horrible	0.27
unwilling	0.24
awful	0.23
useless	0.21
unapproachable	0.21
unprofessional	0.20

Group 5	
passionate	0.42
brilliant	0.42
intelligent	0.38
knowledgeable	0.35
amazing	0.33
incredible	0.32
inspiring	0.31
insightful	0.29
engaging	0.26
interesting	0.25
enthusiastic	0.24
smart	0.24
wonderful	0.23
fantastic	0.22

Group 6	
tough	0.43
hard	0.42
not_bad	0.38
fair	0.31
good	0.29
difficult	0.28
intimidating	0.26
smart	0.25
not_easy	0.25
intelligent	0.23
clear	0.20

Group 7	
hot	0.57
gorgeous	0.49
young	0.42
beautiful	0.40
attractive	0.37
cute	0.33
sexy	0.31
pretty	0.23
cool	0.20