Dimensions of Self-Expression in Facebook Status Updates

Adam D. I. Kramer
Facebook, Inc.
akramer@fb.com

Cindy K. Chung
The University of Texas at Austin
cindyk.chung@mail.utexas.edu

ABSTRACT

We describe the dimensions along which Facebook users tend to express themselves via status updates using the semi-automated text analysis approach, the Meaning Extraction Method (MEM). First, we examined dimensions of self-expression in all status updates from a sample of four million Facebook users from four English-speaking countries (the United States, Canada, the United Kingdom, and Australia) in order to examine how these countries vary in their self-expressions. All four countries showed a basic three-component structure, indicating that the medium is a stronger influence than country characteristics or demographics on how people use Facebook status updates. In each country, people vary in terms of the extent to which they use Informal Speech, share Positive Events, and discuss School in their Facebook status updates. Together, these factors tell us how users differ in their self-expression, and thus illustrate meaningful use cases for the product: Talking about what's going on tends to be positive, and people vary in terms of the extent to which their status updates are short, slangy emotional expressions and topics regarding school. The specific words that define these factors showed subtle differences across countries: The use of profanity indicates fewer school words (but only in Australia), whereas the UK shows greater use of slang terms (rather than profanity) when speaking informally. The MEM also identified English-language dialects as a meaningful dimension along which the countries varied. In sum, beyond simply indicating topicality of posts, this study provides insight into how status updates are used for self-expression. We discuss several theoretical frameworks that could produce these results, and more broadly discuss the generation of theoretical frameworks from wholly empirical data (such as naturalistic Internet speech) using the MEM.

1. INTRODUCTION

1.1 Text Analysis and the MEM

As automated text analysis is growing in popularity, there is also a growing belief that natural language provides a remarkably accurate window into our identities [15,10] and psychological states [24,25]. In fact, there are several text analysis approaches that have been developed specifically for the measurement of words as reflections of personality and psychological states. For example, the Meaning Extraction Method (MEM) is a semi-automated text analysis technique used in the field of personality psychology on personal narratives to extract dimensions along which people think about themselves or a particular topic [7]. Initially developed for the analysis of self-expressive text such as self-descriptive essays, the MEM can be used to understand self-expression in social media. In this paper, we apply the MEM to Facebook status updates to extract dimensions along which people express themselves, and compare dimensions of self-expression in Facebook status updates across the United States (US), Canada, the United Kingdom (UK), and Australia. Finally, we discuss how a bottom-up approach such as the MEM might be used to assess dimensions of self-expression across different social media in a reliable, efficient, unbiased, and unobtrusive manner (c.f. [24,26]).

1.2 Facebook and Status Updates

We applied the MEM to Facebook status updates because short-format "microblogs" such as these have been argued to serve many purposes, ranging from self-expression to news aggregation (i.e., providing links to articles of interest to the presumed audience), to updates regarding a group or organization (e.g., a band announcing a tour or album release), and also vary in terms of the goals and intentions of the user [1,21,28]). Due to the control that Facebook offers over how updates are broadcast (i.e., users can configure any update to be broadcast to the public, to their friends, or to a subset of their friends), Facebook status updates may provide a more authentic source of self-expression [19,28]. Facebook status updates have also been shown to promote psychological well-being via many of the same processes as offline social interaction. For example, [38] showed that viewing one's Facebook page serves the psychological goal of self-affirmation (with effect sizes comparable to offline manipulations) and [5] shows that interacting with others via Facebook improves social connectedness (while also noting that "lurking" or passively consuming without contribution is depressing, much like sitting in a corner at a party).

Status updates provide cues to the psychological state of individual users, and when examined collectively, have been shown to provide insight into the psychological state of groups who update. For example, [24] analyzed the status updates of 400 million Facebook users in America over time. By counting the relative rates of positive and negative emotion word use, [24] was able to clearly identify culturally shared positive events in America (e.g., national holidays such as Thanksgiving, Christmas, Valentine's Day, as well as the most celebrated day of the American work week, Fridays). Through the same word counting index, [24] also identified culturally shared negative events (e.g., the anniversary of the 9/11 attacks, the sudden death of Michael Jackson, the Chilean earthquake of 2009, and the most dreaded day of the American work week, Mondays). Overall, [24] presented an ecologically valid way to assess Gross National Happiness over time using an automated approach to analyze naturalistic Internet text (i.e., Facebook status updates). More broadly, the study showed that Facebook status updates not only reflect the psychology of individuals, but they can also characterize groups, cultures, and countries.

Indeed, the ever-increasing user base of Facebook status updates provides a powerful population to assess the psychology of large groups. Facebook started in 2004 with an initial user base of Harvard University students in Massachusetts. Very quickly, Facebook expanded to other local universities, and eventually to all high school and university students in the United States in 2005. In 2006, Facebook eventually opened to anyone over the age of 13 with an email address [1]. Despite these initial

demographic restrictions, Facebook has since shown growing and remarkable representations of people from other demographics, with over 30% of the user base being over the age of 35 in 2010 and over 20% of the population of countries ranging from Hong Kong to Chile to Iceland being represented on the site [4]. In sum, Facebook status updates offer a huge archive by which to assess self-expressions across cultures, with user bases that are more representative of a population's daily natural language in many countries than most other archival sources.

1.3 Our Goals

These intercultural differences are a primary topic of our paper: Can we identify how citizens of different countries are using Facebook's "status update" product? Are the basic usage patterns of these countries basically the same, or demonstrably different? Can we extract dimensions of self-expression that go beyond simple quantity of sharing and the sharing of emotional content? By understanding what is being expressed in a naturalistic context, we can understand how the medium is being used. In this study, we ask: What are the dimensions along which people express themselves when they update their status on microblogs, and on Facebook in particular, and do these dimensions differ for different countries? With knowledge of such a dimensional structure, we can then see if the medium has been adopted for use differently in different countries. By comparing English-speaking countries only, we can examine whether there are any differences in how Facebook status updates are used to express dimensions of the self without the complexity of cross-language analyses (though we believe that cross-language research in this ilk to be a very interesting avenue for future study).

2. THE MEANING EXTRACTION METHOD EXPLAINED

The MEM [7] relies upon the "lexical hypothesis," which posits that important concepts become represented as single lexical items (words) in that language [11]. When applied to personality, important concepts that describe a person's character become represented as a single word in that language. The lexical hypothesis has been used to derive one of the most extensively and empirically supported personality structures in the field of Psychology, the "Big Five" personality structure [11].

The MEM begins with the same "lexical hypothesis," but unlike the Big Five (which was derived using Likert ratings of adjectives on self-report scales), the MEM examines the co-occurrences of words within natural language-based self-descriptions or any text-based corpora, constituting an inductive, and potentially unobtrusive approach [26]. This approach extracts a set of lexical items, which together represent a relevant dimension of word usage in the corpus; if users who use certain words also tend to use other words, then the set of words, which are co-used together, is argued to represent a meaningful quality of the corpus [7,26]. From psychological studies in which participants have been asked to provide personality self-descriptions, the MEM produces common dimensions along which people tend to think about themselves. For self-descriptions in Spanish (by Mexican college students), culture-specific dimensions not found in comparable American samples tend to emerge, such as Simpatía, which includes word co-occurrences such as *affectionate, honest, noble,* and *tolerant* [31]. The MEM, then, is a culturally unbiased method to derive meaningful dimensions along which people vary.

The MEM has also been used to examine regional differences in values across America. In a corpus of "This I Believe" essays (essays on the beliefs and values that guide people's lives), various dimensions of values, such as religion, health, and community emerged [8]. By correlating the component scores of the value themes with state-level indicators, [8] found that the religion theme (e.g., *god, church, Christian*) was more likely to be mentioned in states that had a higher proportion of children who attended religious services weekly; the health theme (e.g., *hospital, doctor, cancer*) was more likely to be mentioned in states with the highest rates of death due to chronic illness; the community theme (e.g., *friend, meet, town*) was more likely to be mentioned in states with the greatest number of restaurants and movie theatres per capita. While these results may not be surprising, they serve to illustrate the deep influence that our locale, culture, and upbringing have on the way we choose and use words. In summary, by using the MEM, researchers were able to capture valid dimensions of values that reflected regional variations in practices, illness rates, and institutions.

3. THE PRESENT STUDY

In this paper, we used the MEM to derive dimensions of self-expression from Facebook status updates. In order to determine any differences in self-expression across countries, the corpus of words is composed of all status updates made by four million randomly selected Facebook users from four of the largest countries with a primarily English-speaking user-base: Australia, Canada, the US, and the UK. The MEM is an alternative to "keyword extraction" computational methods such as latent semantic analysis (LSA, [22]): While LSA, Latent Dirichlet Allocation (LDA, [3]), and computer-reading approaches abound with the stated goal of extracting meaning from natural text [22], the word-count approach, used more frequently in psychology [20,29,31] offers several advantages for our goal: the MEM approach allows for the extraction of linguistic dimensions rather than keywords, which then can be examined at the level of the individual (i.e., we can quantify where someone lies on a dimension). These linguistic dimensions are also technologically more efficient to compute: Words can be parsed and counted on a row-by-row basis and then aggregated in one pass (i.e., correlated), allowing for massively parallel computation [24,37]. This, in turn, allows for the examination of a much larger set of words and a more representative set of users. Other approaches such as WordNet or PMI-IR may provide other insights regarding users' behavior, but these methods lack a straightforward manner by which to simultaneously examine word usage both within and across users (e.g., [22]). While other methods such as LDA may potentially produce similar results, the MEM procedure is much more direct and interpretable as it is based on interpretations of principle components analysis rather than an iterative multi-stage latent variable model. In other words, it can be completed and understood by social scientists with specific interest in the result as well as computational modelers with specific interest in the method; further, as the MEM has been shown to reveal dimensions of self-expression in personal narratives with sensitivity to detect regional and cultural differences in self-expression [7,8], we chose the MEM as an ideal, bottom-up method to identify dimensions of self-expression in Facebook status updates across countries.

3.1 Method

All status updates from four million Facebook users were analyzed: One million users who self-reported being from each country on their Facebook profile. The random selection and anonymization of users, along with the text processing of all status updates was conducted using the Hive database framework on the Hadoop MapReduce framework [37]. Hive allowed use of the

TAWC word-counting software [25], allowing for researchers to avoid viewing any analyzed user's updates, in keeping with the Facebook Terms of Service [10]. For each user, all updates (if any) made between September of 2007 and February of 2010 were analyzed; this long timeframe ensured that a reasonable amount of text produced by each user was represented in the analyses. We first established that our sample was diverse in terms of age (i.e., not comprised entirely of college students): sampled users had a mean age ranging from 33.3 and 35.4 years across countries with standard deviations between 13.8 and 14.6 years. These statistics indicate that more than simply "college students" are represented on Facebook (as indicated by the high mean), and that a wide age range is represented in our data (as indicated by the rather high standard deviation)[1].

	USA	CAN	UK	AUS
# Users	437,370	466,411	450,061	411,982
Age in Years (SD)	32.8 (14.6)	31.7 (13.9)	30.4 (13.1)	30.7 (13.5)
Percent Female	58.5%	56.4%	53.6%	56.4%
Mean # Status Updates per User (SD)	138.4 (210.6)	187.9 (259.9)	162.7 (232.8)	120.5 (168.5)
Mean # Total Words Across All Status Updates by a User (SD)	1650 (2784.7)	1939 (3136.9)	1766 (2799.5)	1307 (2032.9)

Table 1. Final sample of Facebook users from each country. Note: The table includes the subsample of Facebook users who wrote at least 100 words in their history of status updates.[1]

First, a list of single lexical items (i.e., individual words) was ranked in terms of how frequently they were used across the entire corpus. This list was then filtered to remove non-content words (for example, function words such as *and, the, was, you,* etc.). The remaining words were re-ranked by the proportion of sampled users who had used that word in any status update. After this point, users whose histories of all status updates contained fewer than 100 words total were dropped from further analysis, leaving 1,765,824 total users (see Table 1 for breakdown by country).

From this ranked list, the top 1,000 words in terms of occurrence (a word was ranked higher if it occurred more frequently in the corpus, as in [7]) were converted into the custom dictionary format specified in the popular word-counting system LIWC [29]. Specifically, each word was stemmed and formatted using letter patterns or regular expressions to capture all potential uses of the word in a single category. For example, the presence of the string "bbq" in the top 1,000 words resulted in a single category including the stem with alternate spellings "bbq*" (encompassing "bbq" and "bbqing"), as well as "barbq*", "barbeq*", and others. This procedure resulted in 760 separate word categories, which represented 8% of all words used in the set of status updates and 97% of users who had used words from one of the categories[2]. This custom LIWC dictionary was created by hand and is available from the authors upon request.

Next, the corpus was examined to determine whether each user had ever used a word from each word category. This resulted in a 1,765,824-by-760 user-by-word matrix, each row of which contained either a zero or one, representing whether the user used the lexical item in question in one of their status updates. These items were then correlated across users, producing a 760-by-760 correlation matrix for the total sample, and one for each country. No correlations were significantly lower than zero, and the largest observed correlation was 0.65; this pattern is common in analyses such as these due to the fact that people who use more words use a wider variety of words [26]). This correlation matrix was then subjected to principal components analysis (PCA) with Varimax rotation, as in [7,26,31]. The loadings of this matrix are the product of the MEM: They represent the extent to which use of one word in any of a user's status updates predicts use of other words in general from the respective component.

	USA	CAN	UK	AUS
1	day	day	day	day
2	loud	hangover	loud	loud
3	word	loud	long	ticket
4	ticket	ticket	word	word
5	nice	word	ticket	good
6	long	text	nice	light
7	light	light	hangover	hangover
8	hangover	winter	light	text
9	good	good	good	nice
10	vote	vote	vote	lie

Table 2. Top 10 most commonly used words (excluding function words) in Facebook status updates for each country. Note: Frequency was determined as the proportion of users who used a word.

To maximize what can be learned from Facebook status updates and interpreted via the MEM, and in order to promote and/or assist other researchers in using MEM approaches for their own data sets, we have published the aggregate matrix as well as each country's matrix in the auxiliary materials.

3.2 Results and discussion

3.2.1 Aggregate Matrix

In order to describe how Facebook status updates are used overall, we first conducted an analysis using an aggregate matrix of all four English-speaking countries. The results from this aggregate matrix produced one large "wordiness" component, which contained positive loadings for 732 (96%) of the word categories. Users' scores on this component were correlated with the number

[1] Due to the number of subjects, these differences are significant but arguably meaningless due to negligible effect sizes; all *d*'s < 0.1.

[2] The remaining 3% were not qualitatively examined as such an examination could constitute an invasion of users' privacy [10].

of words used in the user's set of status updates, $r = .58$, and with the number of updates they had made, $r = .62$. The same result has been found for blogs [11], and is consistent with the fact that people who talk more in their everyday lives tend to use a greater range of words.[3] We followed the approach of [26], and removed this first component prior to rotation.

Following the procedure described in [26], five components were extracted. We considered a loading meaningful if it had an absolute value greater than 0.30. The solution for the aggregate matrix had simple structure (each lexical item loaded on to at most one component). The component with the largest loadings contained negative loadings for Britishisms such as *arse*, *pub, mate,* and *England*, and positive loadings for Americanisms such as *mom*, (*mum* is used in British English), *laundry* (as opposed to *the wash*), *vacation* (as opposed to *holiday*), and *halloween* (a secular holiday celebrated primarily in the US and Canada, much less so in the UK, and hardly at all in Australia). This dimension was bimodally distributed, with those from the UK and Australia showing higher use and those from the US and Canada showing less use of Britishisms. This finding provides evidence of the efficacy of the method: With text from groups of people who speak different dialects, these dialects appear in the components, suggesting that the components reflect meaningful features of the users. This large first factor supports that the method is "tuned in" to the most relevant linguistic features of the corpus along which people tend to vary and extracts the most notable dimensions.

Comp 1	Comp 2	Comp 3	Comp 4	Comp 5
Weekend	Hubby	Mom (R)	Lol	Day
	Daughter	Arse		Time
	Essay (R)	Pub		Good
	Bless	Mate		Likes
	Exam (R)	Vacation		Love

Table 3. Top five highest-loading words (all load with absolute value above 0.3), by component, for the aggregate matrix.

Two of the four additional components included only one word each (*weekend* and *lol*, respectively), one contained positive loadings for family-oriented terms (e.g., *hubby*, *daughter*, *bless*) and negative loadings for school-oriented terms (e.g., *essay exam*), and the last contained positive loadings for both time words (e.g., *day*, *time*) and positive emotion words (e.g., *good, like, love).*

3.2.2 Individual Country Matrices: The United States

To examine whether the countries differed in their status-update dimensions, their individual correlation matrices (transformed to *Z*–score matrices) were compared using a random effects model: After removing a random effect for each of the two word categories (effectively the row and column category), we tested whether knowing which country generated a correlation could explain significant variance among correlations. This analysis was conducted using the *lme4* package for the statistical program R [30]. Though there was a large degree of commonality among countries as well (Cronbach's α among countries = 0.97), due to our large sample size we found a significant difference among countries in the pattern of correlations, $\chi^2(3) = 40{,}202$, $p < .001$. As such, we examined the four countries separately with an eye towards describing both similarities and differences.

The US was examined first as Facebook was initially created in the US, making these users a part of Facebook's largest and most senior user base. Examination of the correlation matrix suggested either a two-component or five-component solution: Components 1 and 2 were the same in both the two-component and the five-component solution (see Tables 3 and 4): One component contained profane or "slangy" words, including Internet neologisms such as *wtf* and *haha.* Upon comparison to the aggregate matrix, this factor generally resembled the aggregate matrix component represented by the term *lol* in terms of the ordering of words on this component (though the aggregate matrix did not show as many of these items loading above our absolute value cutoff of 0.30). This component was labeled "Informal Speech." In other words, the extent to which Americans speak formally or informally is a primary means by which they can be differentiated using the MEM; variability in the corpus of American status updates can be explained in terms of how formal the speech is. This is consistent with [26], who showed that the use of "ranty" words (including profanity) was a differentiator of weblog authors.

The second-largest component (in terms of variance explained) also contained positive loadings for both "time" terms and "positive affect terms". This component, labeled "Positive Events," resembled the final component found in the aggregate matrix. In general, this component represented the extent to which people discussed their daily activities.

This category, at first glance, may appear to represent a single dimension with no clear poles: Positive words are believed to represent the psychological "emotion" construct. Interestingly, no negative words were loaded on this factor below -0.30. We believe that the Positive Events factor effectively indicates a medium use dimension. In other words, Facebook seems to be a site to which people go to share positive events in their life, or simply to share events in their life when they are feeling positive. The absence of negative loadings for negative words, then, indicates that negative words are used in a manner orthogonal to discussion of time or events. Note also that words about being solitary are not used in the sharing of positive emotions.

There is some support for this interpretation in the psychology literature. [35] describe and briefly review evidence that those who are more extraverted (oriented towards social interaction) tend to be happier, more positive, and more satisfied with life. The research in [35] also provides a duality that nicely dovetails with the "Positive Event" factor: First, people who are more extraverted tend to be happier in general, such that those who engage in more events will both be happier and have more events to talk about. Second, all people (even introverts) who actually engage in social activities tend to express increased positive affect during and following these events, such that those who engage in events and share them with friends may be more positive in general. Finally, we note that people who are more extraverted in general may be more likely to share information on social networking sites in the first place (even if there is evidence that the Facebook population is not abnormally extraverted [5]): If these people are more positive as [35] show, greater sharing could lead to this factor.

Three other potential explanations could also have created this factor: First, the timescale of negative emotionality is longer.

[3] We also note that for corpora made up of texts with few words (our minimum was 100 words) it is unreasonable to expect a person's full vocabulary to be represented.

Because negative events engender rumination prior to expression [32], they may not be sufficiently formed when the status update is made (because the status is being updated during or shortly after the event, or it might seem irrelevant). Second, there is evidence that people are simply more hesitant to express negative emotionality: Negative emotions may be more private [32], which may drive negative words to be too rarely expressed in to explain meaningful amounts of variance. Finally, there is growing evidence that happiness is, in general, driven by experiences (especially social experiences) rather than by factors such as possession of wealth or material goods [9]; thus, positive emotion expression may be *driven* by experiences: When people are naturalistically self-motivated to report on what has made them happy, it is actually their experiences that rise to the top.

As with the Informal Speech component, it is important to note here that the component itself does not represent a dichotomy so much as a dimension along which it is useful to differentiate Facebook users. For example, it is not the case that a given person is either an introvert or an extrovert (though researchers have been known to falsely dichotomize extraversion); one could be anywhere in between, just as users can score anywhere along the positive event scale (even in the middle).

The five-component solution also produced a third component with school-related words such as *homework, study, essay,* and *English.* This component resembled the aggregate matrix's third component, and was labeled "School," indicating that words about schoolwork tend to be highly correlated. This does not, however, mean that most American Facebook users are in school—if everybody was in school, everybody might use these so-called "school" words, which could produce a *low* correlation as correlations are measures of "shared" variance and there is very little variance to begin with. Rather, this factor indicates that people who use *some* "school" words tend to use others, whereas other people rarely use any of them—in other words, the relevance of school as a thing to post about is a dimension on which people vary. Prior research using the MEM on college students alone also shows this factor [7,31], emphasizing that focus on school is indeed a primary dimension among which even college students vary. So, while school-related words tend to cluster together, this does not necessarily indicate that the user base is primarily composed of students, but rather that people who use some school words tend to use others (i.e., some people—probably students—tend to think along a school dimension).

The fourth and fifth components, after rotation, explained very little variance of the overall matrix (about 1% each); the fourth contained only one loading (*trust*), and the fifth contained no loadings that were greater in magnitude than 0.30. As such, these components are left to future study.

3.2.3 Individual Country Matrices: Canada, the United Kingdom, and Australia

The correlation matrices for the three other English-speaking countries were also examined using PCA. While different when analyzed in aggregate, the component structures were nearly identical when analyzed separately: When a five-component structure was extracted, the same three components appeared: Informal Speech, Positive Events, and School (Table 4). This is possible because in the aggregate matrix, British English words like *pub* and *mate* were correlated because half of the population (those from the US and Canada) used these rarely if ever, whereas the other half (those from the UK and Australia) used them consistently. As such, when PCAs were computed within each country, these words did not form a factor.

	Informal Speech	Positive Events	School
US	Fucked, shitty, bitch, ass, haha	Day, time, good, love, likes	Homework, study, essay, English, exam
UK	Fucked, shitty, lol, sum, haha	Day, time, love, look, happy	Essay, exam, hubby (R), daughter (R), gay
CA	Fucked, shitty, bitch, gay, ass	Day, time, happy, love, good	Study, exam, essay, homework, English
AU	Hubby (R), daughter (R), fucked, gay, wonderful (R)	Day, time, love, night, good	Exam, study, essay, uni, fucked (R)

Table 4. Top five words for each country for each of the three replicated components.

To statistically examine the extent to which these components represented the same construct, the loading matrices for the four countries were examined (i.e., the loadings for Informal Speech generated from the US data was correlated with the loadings for Informal Speech for Canada, the UK, and Australia). Out of the twelve correlations (three components times four countries), the lowest was $r = .79$, again, supporting large similarities across countries. Qualitatively, one notable difference was the School factor for the UK and Australia versus the US and Canada: These components included negative loadings for family-oriented words (e.g., *hubby, daughter*), suggesting that the differentiation between those attending and not attending school in the UK and Australia may fall more along the lines of a "School *versus* Family" dimension than simply "School or not."

Australia also stands out as the only country to show profanity loading onto more than just the "Informal Speech" category, showing up as a negatively loaded word category on the "School" factor: Those who use more school-related words are less likely to use *fuck* in any of its forms, consistent with Australia's reputation as a country that is historically linguistically profane (e.g., [39], but see also [36]). Australia also shows the words *day* and *night* as two of the top four highest-loading words on the "Positive Events" factor (and both positive), further indicating that the factor is more about time in general than it is about specific times. In other words, some people tend to think along a time dimension or not, rather than about a specific time of day [7].

Beyond a qualitative read of the factors to describe differences among countries in how status updates were used, the scale of our data allowed for a statistical comparison of word loadings. To see which words loaded onto certain components for only one country (and thus differentiated one country from others), bootstrapping techniques were used to estimate the standard error of deviations from our loading magnitude cutoff, 0.30. This was then used to calculate whether a given loading magnitude for a given component in a given country was significantly different from 0.30. We then examined words that were significantly above the cutoff magnitude in some countries but significantly below it in others (See Table 5).

More than highlighting differences, Table 5 shows that there is indeed remarkable *consistency* in terms of which words load on to which factors. However some cross-cultural instances also appear: The Britishism *hubby* is used by Canadians as a negative marker of informal speech; those who use this affectionate term for "husband" tend to use fewer informal speech cues—but only in Canada. The same goes for the term *wonderful*, whereas the term *gay* in Canada is an indicator of less formal speech. In other countries, these words do not indicate more or less informal speech. This could be because Canada, a member of the Commonwealth, may have some citizens speaking in a more UK-consistent manner, or perhaps Canadians use the word *gay* more as a slur ("that's so gay") than as a descriptor of sexual orientation or as an emotion word. Conversely, in the UK (where people are considered to be more formal in general), more slang terms (as opposed to profanity) loaded on to the Informal Speech category. Future research could address and test these hypotheses directly.

	Informal Speech	Positive Events	School
USA	None	None	None
CAN	Hubby (R), Wonderful (R), Gay	Wait, Wishes	None
UK	lmao, sum, lol	None	None
AUS	None	None	None

Table 5. Category-defining words for only one country.

4. GENERAL DISCUSSION

4.1 Summary

Across four English-language speaking countries, we found three components, Positive Events, Informal Speech, and School, along with cultural differences in medium use: In the UK and Australia, the opposite pole from the school–oriented words corresponded to family-oriented words. This method also bore out several stereotypes about these countries: Australians' reputation as being a more "wild" culture is borne out as a broader incidence of profanity (i.e., *fuck* loading on to multiple factors), whereas the reputation of the British as being "reserved" was shown in the greater incidence of slang terms (rather than profanity) in the factor representing formality of speech. These countries were not entirely different, however: Both used a set of "British English" words that are relatively rare in the United States and Canada, as shown in the decomposition of the aggregate correlation matrix. This duality, of both contrasting countries and showing how they are similar, is a feature of the MEM method.

Beyond simply indicating topicality of posts, this study provides insight into how status updates are used for self-expression: Talking positively about what's going on is a meaningful dimension along which users of the system vary: Talking about what's going on tends to be positive. People also vary in the extent to which their status updates are short, slangy emotional expressions and topics regarding school. The MEM also illustrates the diversity of the Facebook user base: The MEM was able to identify English-language dialects as a meaningful dimension along which the aggregate sample varied, and the School dimension as a meaningful dimension along which the individual countries varied.

Indeed, these findings regarding common uses of the medium are even stronger than the differences among countries. Few differences were found among the countries in terms of their use of language in status updates. This could be for several reasons: First, perhaps these countries simply do not differ much in terms of language or use of the medium. Facebook was started in the United States; social norms developed there might serve as models for people in other countries once Facebook.com was opened to them. [6] showed that social modeling of Facebook behavior by newcomers is indeed quite strong. That these emotions and norms in general are transmitted through and modeled in terms of text-only interactions procedures is also well established (e.g., [14,16,18,23]). Another possibility is that the affordances of the medium effectively overpower differences in self-expression that exist across the four countries. This would be a strong effect of the medium, but also of the countries themselves, which share a language and a common cultural heritage. Examination of native English-speaking Facebook users who did not grow up in, say, the United Kingdom (for example, children born to citizens of the UK but raised abroad) could address the extent to which these similarities are due to the medium or the culture. Replication of the research described herein, using words from a different language, would also address this concern and provide an interesting avenue for future research.

4.2 Theoretical Implications

The MEM can be used as a tool for both generating and testing theories of language within a medium. As the "corpus" of Internet text grows by terabytes per day, the questions of what linguistic behavior and corresponding psychological information we should expect to see and learn from a certain subcorpus also grows. While many theoretical frameworks exist to generate hypotheses about Internet speech, the MEM provides an opportunity for researchers who are dogmatically neutral with regard to word use to effectively generate theoretical frameworks empirically [17]. Without a theoretical background, three dimensions of status update use were generated and effectively replicated across four different English-speaking countries: The question, then, is what to make of these—what theoretical framework would drive a three-component structure of this sort? Would these factors replicate for different CMC media? One theory could be based on the expected audience of users: Facebook allows users fine-grained control over who can see a given status update: Friends only, friends of friends, specific friends, or the whole world, which would drive users to share items they believe to be of interest to their target audience, either because they share many qualities of their life (i.e., they share the "student" vocation, generating the "School" component), or because they believe their audience to be interested in the high points of their days (thus generating the "Positive Events" factor), or allowing them to talk in a more casual way than they would to non-friends. This theoretical framework would then expect different factors to appear for more public media such as blogs or tweets.

Another theoretical framework involves the undirected nature of status updates: These updates are "broadcast" to the friends of the author, rather than "directed" at one or more friends. This could lead to more author-centric text (descriptions of day-to-day life, including schoolwork for students and events) whereas a directed context (such as wall posts or emails) may not have generated these factors. The length of status updates (up to 420 characters, but usually far fewer) could also encourage greater use of common ground between the author and the audience, indicating that these components would more directly replicate in other short-format media (such as tweets or text messages) rather than long-format media (such as blogs). The theoretical frameworks and predictions above are made possible by the MEM when viewed as a "bottom-

up" theory generation framework: By observing large-scale naturalistic data and aggregating them in a forthright manner, we are able to see what the data show and to form theories of "why" accordingly, which can then be tested via parallel analyses.

4.3 Future Directions

Although the MEM is able to summarize dimensions of self-expression for millions of users and billions of words, this unobtrusive method is not a trivial undertaking: Stemming and filtering non-content words is currently not automated. Similarly, the method itself relies upon the interpretation and labeling of PCA components: Without hypothesis-driven research (such as testing whether students in school use more School words than students out of school or non-students), it may be premature to call a component a "School" component for any purpose beyond ease of reference (see [34] for discussion on how to name components).

Future work could investigate why users tend to be higher or lower on one of these dimensions. For example, the "School" interpretation of the School component could be validated by showing that current students score significantly higher on this component than non-students, by showing that users with social networks that have a high proportion of shared school affiliations score higher on this component, or by examining the words used to describe day-to-day activities: People who go to school may use a very similar nomenclature regardless of their location (i.e., they describe their activities in the same manner regardless of which school they attend or which country they attend school in), while those not in school may not use words in a pattern that is detectable at the gross national level (i.e., they may have a job in which there are too few employees to count as more than "noise").

Future research could seek to predict which people (extraverts versus introverts, older versus younger users, etc.) are likely to post about Positive Events, or use Informal Speech. In effect, the MEM can help researchers to identify the fundamental ways in which word use differs, which in turn can help identify more relevant variables to explore: For example, do people talking about positive events generate more or fewer wall posts and comments? Do users who use fewer school-related words (or as users decrease in use of these words), does the content or manner in which they post change? Does informal language use predict possession of more or less social capital? Are particular types of status updates associated with affiliation with a greater number of groups or with particular types of groups? Future research can compare demographic groups on factors defined by the whole set, or examine whether the set of MEM components itself replicates across groups (in a manner akin to a confirmatory factor analysis; [12]). In short, once quantified (i.e., by examining loading scores for actual text), the ways in which status updates are used for self-expression can be used as signals of other communication, affiliation, and network utilization practices.

Comparisons among communications media may also distinguish how certain media are used: Is positive event description higher for directed communication (e.g., public Facebook wall posts or private emails), less directed longer-format communication (e.g., blog posts), or anonymous communication (such as long-format undirected diaries, public undirected tweets, or public undirected forum posts not connected to a "real" identity)? Does this factor replicate in offline media? Just as some posting behaviors and usage patterns can be localized to specific media and user clusters, these may also indicate user behavior (e.g., likelihood of clicking on advertisements [22]), or other individual differences such as extraversion. Similarly, using the MEM to compare posts on different topics but within mediums (e.g., news-oriented blogs versus diary-style blogs) may enable a better understanding of the user base as well as the ways that different users make use of the same medium.

Finally, we expect that the means by which people express themselves using status updates will change over time. Our current analysis examined all text written by users, which means that our results are better interpreted as an examination of the way that people have expressed themselves using status updates for the past several years; how they are expressing themselves today, or in the coming years is independently interesting.

5. CONCLUSION

We applied a method developed in the Personality Psychology literature, the Meaning Extraction Method, to extract meaningful dimensions of self-expression in Facebook status updates. Users across four different English-language speaking countries used the product in a very similar manner, varying primarily in terms of the use of informal language, sharing positive events, and focusing on school. Our findings were also used to explore differences among four English-speaking countries. We found some country-level differences regarding formality of speech, but the more remarkable finding was that the four countries showed remarkably similar components.

The MEM can be used to identify dimensions along which users differ both within and across social media or between user types. The raw MEM scores for these components can be used in future research in order to determine user characteristics for people who post in these ways. The analyses tell us how a medium is being used, and gives hints as to how social media is shaping communication (e.g., addressing what kinds of expression are encouraged) and to how communication might shape social media (e.g., development of forums for specific kinds of expression), as discussed above.

6. ACKNOWLEDGMENTS

The second author was supported, in part, by funding from the Army Research Institute (W91WAW-07-C-0029) and DIA (HHM-402-10-C-0100). We thank the Facebook Data/Science team for their support, and also acknowledge James W. Pennebaker and Moira Burke for their comments on an earlier draft of this paper.

7. REFERENCES

1. Abram, C. (2006). Welcome to Facebook, everyone. *The Facebook Blog*, retrieved Februray 1, 2011. http://blog.facebook.com/blog.php?post=2210227130
2. Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality not self-idealization. *Psychological Science, 21*, 372-374.
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3,* 993-1022.
4. Burbary, K. (2010). Dispelling the youth myth: Five useful Facebook demographic statistics. *Personal weblog.* Retrieved January 26, 2011 from http://www.kenburbary.com/2010/01/dispelling-the-youth-myth-five-useful-facebook-demographic-statistics/
5. Burke, M., Marlow, C., & Lento, T. (2010). Social network activity and social well-being. *Proc. CHI 2010,* 1909-1912.

6. Burke, M., Marlow, C., & Lento, T. (2009). Feed me: Motivating newcomer contribution in social network sites. *Proc. CHI 2009,* 945-954.
7. Chung, C. K., & Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality, 42,* 96-132.
8. Chung, C. K., Rentfrow, P. R., & Pennebaker, J. W. (January, 2011). *Mapping beliefs across America: Validity of themes in "This I Believe" essays*. Talk presented at the 2011 Annual Meeting for the Society of Personality and Social Psychology, San Antonio, TX.
9. Diener, E. Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist, 55,* 34-43.
10. Facebook Terms of Service (2010). Retrieved from http://www.facebook.com/terms.php on February 6, 2011.
11. Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology, 59,* 1216-1229.
12. Gorsuch, R. L. (1983). Factor Analysis (2nd Edition). Hillsdale, NJ: Lawrence Erlbaum Associates.
13. Grace, J. H., Zhao, D., & Boyd, D. (2010). Microblogging: What and how can we learn from it? *Proc. CHI 2010,* 4517-4520.
14. Grice, H. P. (1979). Logic and Conversation. In P. Cole and J. Morgan (Eds.), *Syntax and Semantics 3: Speech Acts,* pp. 26-40. New York: Academic Press.
15. Gonzales, A. L., & Hancock, J. T. (2008). Identity shift in computer-mediated environments. *Media Psychology, 11,* 167-185.
16. Guillory, J., Spiegel, J., Drislane, M., Weiss, B., Donner, W., & Hancock, J. T. (in press). Angry now?: Emotion contagion in distributed groups. *Proc. CHI 2011.*
17. Hancock, J. T. (January, 2011). *What lies beneath: Using language to understand deception.* Talk presented at the 2011 Annual Meeting for the Society of Personality and Social Psychology, San Antonio, TX.
18. Hancock, J. T., Landrigan, C., & Silver, C. (2007). Expressing emotion in text-based communication. *Proc. CHI 2007,*
19. Hollenbaugh, E. E. (2010). Personal journal bloggers: Profiles of disclosiveness. *Computers in Human Behavior, 26,* 1657-1666.
20. Ireland, M., & Pennebaker, J. W. (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology, 99,* 549-571.
21. Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. *Proc. SNA-KDD 2007.*
22. Kaur, I. & Hornoff, A. J. (2005). A comparison of LSA, WordNet and PMI-IR for predicting user click behavior. *Proc. CHI 2005*, 51-60.
23. Kramer, A. D. I. (2011). Emotion expression and contagion online: Statuses, sentiment, and sympathy. Poster presented at the 2011 Annual Meeting for the Society of Personality and Social Psychology, San Antonio, TX.
24. Kramer, A. D. I. (2010). An unobtrusive model of "Gross National Happiness." *Proc. CHI 2010,* 287-290.
25. Kramer, A. D. I., Fussell, S. R., & Setlock, L. D. (2004). Text analysis as a tool for analyzing conversation in online support groups. *Proc. CHI 2004*, 1485-1489.
26. Kramer, A. D. I., & Rodden, K. (2008). Word usage and posting behavior: Modeling blogs with unobtrusive data collection methods. *Proc. CHI 2008,* 1125-1129.
27. Mehl, M, Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science, 317,* 82.
28. Naaman, M., Boase, J., & Lai, C.-H. (2010). Is it really about me? Message content in social awareness streams. *Proc. CSCW, 2010*.
29. Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic Inquiry and Word Count (LIWC2007).* Austin, TX: http://www.liwc.net.
30. R Development Core Team (2010). R: A language and environment for statistical computing, ISBN 3-900051- 07-0, URL http://www.R-project.org.
31. Ramírez-Esparza, N., Chung, C. K., Sierra-Otero, G., & Pennebaker, J. W. (in press). Cross-cultural constructions of self-schema: Americans and Mexicans. *Journal of Cross-Cultural Psychology.*
32. Rimé, B. (2007). The social sharing of emotion as an interface between individual and collective processes in the construction of emotional climates. *Journal of Social Issues, 63*, 307-322.
33. Rimé, B. (1995). Mental rumination, social sharing, and the recovery from emotional exposure. In J. W. Pennebaker (Ed.), *Emotion, disclosure, & health* (pp. 271-291). Washington, DC, US: American Psychological Association.
34. Saucier, G. (2000). Isms and the structure of social attitudes. *Journal of Personality and Social Psychology, 78,* 366-385.
35. Srivastava, S. K., Angelo, K. M., & Vallereux, S. R. (2008). Extraversion and positive affect: A day reconstruction study of person-environment transactions. *Journal of Research in Personality, 42,* 1613-1618.
36. Thelwall, M. (2008). Fk yea I swear: Cursing and gender in MySpace. *Corpora, 3,* 83-107.
37. Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P., & Murthy, R. (2009). Hive – A warehousing solution over a map-reduce framework. *Proc. VLDB* 2009, 1626-1629.
38. Toma, C. L. (2010). Affirming the self through online profiles: Beneficial effects of social networking sites. *Proc. CHI 2010,* 1749-1752.
39. Wierzbicka, A. (2002). Australian cultural scripts—*bloody* revisited. *Journal of Pragmatics, 34,* 1167-1209.