Micro-serendipity: Meaningful Coincidences in Everyday Life Shared on Twitter

Toine Bogers
Royal School of Library and Information Science, Copenhagen, Denmark
tb@iva.dk

Lennart Björneborn
Royal School of Library and Information Science Copenhagen, Denmark
lb@iva.dk

Abstract

In this paper we present work on *micro-serendipity*: investigating everyday contexts, conditions, and attributes of serendipity as shared on Twitter. In contrast to related work, we deliberately omit a preset definition of serendipity to allow for the inclusion of micro-occurrences of what people *themselves* consider as meaningful coincidences in everyday life. We find that different people have different thresholds for what they consider serendipitous, revealing a *serendipity continuum*. We propose a distinction between *background serendipity* (or 'traditional' serendipity) and *foreground serendipity* (or 'synchronicity', unexpectedly finding something meaningful related to foreground interests). Our study confirms the presence of three key serendipity elements of *unexpectedness*, *insight* and *value* (Makri & Blandford, 2012), and suggests a fourth element, *preoccupation* (foreground problem/interest), which covers synchronicity. Finally, we find that a combination of features based on word usage, POS categories, and hashtag usage show promise in automatically identifying tweets about serendipitous occurrences.

Keywords: serendipity, Twitter, information behavior, information sharing, everyday life

Introduction

Serendipity has traditionally been defined as the accidental yet beneficial discovery of something one was not looking for directly, and has played an important role in many important scientific discoveries, such as x-rays and penicillin (e.g., De Rond & Morley, 2010; Merton & Barber, 2004; Van Andel, 1994). Serendipity also plays an integral part in everyday information behavior when "chance encounters with information, objects, or people [...] lead to fortuitous outcomes" (Rubin et al., 2011). As a consequence, methods and technologies for stimulating and supporting serendipity have received much attention in the field of information science. Technologies such as search engines, micro-blogging, and recommender systems have all been suggested as possible tools for increasing the potential for serendipity (see, e.g., McCay-Peet & Toms, 2011; Piao & Whittle, 2011; Zhang et al., 2012).

However, neither the study of the phenomenon serendipity nor the use of the concept in information science are without their difficulties. One problem is that there does not appear to be a single agreed-upon definition of serendipity: different definitions focus on different aspects, such as whether a serendipitous finding can be related to the active (foreground) information seeking task, or whether it has to be related to the background task alone. Different definitions assign different weights to personal and environmental factors. People also experience serendipity differently, have different thresholds for calling something serendipitous, and may use it synonymously with synchronicity, diversity, or novelty. What is needed, is a better understanding of the different ways people experience and communicate serendipitous occurrences in everyday life.

Acknowledgements: We thank Victoria Rubin, Christina Lioma, Rune Rasmussen, Louis Jensen, and Erik Frøkjær for their valuable help and advice. We also would like to thank tweeters for sharing their tweets.

Bogers, T., & Björneborn, L. (2013). Micro-serendipity: Meaningful coincidences in everyday life shared on Twitter. *iConference 2013 Proceedings* (pp. 196-208). doi:10.9776/13175

To address these issues, Erdelez (2004) called for approaches based on data generated by participants themselves to provide for more naturalistic studies of everyday serendipity. This was echoed by Rubin et al. (2011) who argued that most previous studies have been based on respondents' descriptions elicited in interviews with researchers. They addressed this themselves by analyzing how bloggers describe their everyday serendipitous experiences (op.cit.). We follow up on their approach in this paper by using non-elicited, self-motivated user data from Twitter, the world's largest online micro-blogging platform, to analyze how users share serendipitous experiences in the context of everyday life. We refer to this type of everyday serendipity as *micro-serendipity*.

Whereas Rubin et al. (2011) excluded blog posts that did not contain "a rich description including a mention of an accidental find and a fortuitous outcome", we did not select user experiences in the present study based on a preset and fixed definition of serendipity. Instead, we want to understand what users *themselves* consider as serendipitous experiences and how they actually describe these experiences. We therefore deliberately omit a preset definition of serendipity in order to allow for the inclusion of micro-occurrences of what people themselves consider as meaningful coincidences in their everyday life.

With this research setting in mind, we address the following three research questions, which are part of an ongoing research project dealing with investigating everyday contexts, conditions, and attributes of serendipity as communicated on Twitter:

- What types of serendipity do Twitter users experience and communicate using Twitter? (**RQ1**)
- How often do people share serendipitous experiences on Twitter, and are there large individual differences in its frequency? (**RQ2**)
- What terminology do people use to describe and share serendipitous experiences on Twitter? (**RQ3**)

In sum, we see our main contribution as a detailed analysis of Twitter as a source for research into micro-serendipity as outlined above.

The remainder of this paper is organized as follows. The next section discusses our methodology with regard to data collection and annotation. The subsequent three sections ('Experiencing serendipity'; 'Frequency of serendipity'; 'Describing serendipity') address our research questions in turn, including both quantitative and qualitative analyses for each question. We discuss our findings and present our conclusions in the last section, along with plans for future work. Related work is discussed where directly relevant.

Methodology

Twitter has become a popular data source for social science research (e.g., Thelwall et al., 2011) with over 340 million tweets per day[1] providing an unprecedented window into everyday life experiences, thoughts, interests, conversations, and language use of hundreds of millions of people. Our goal is to examine whether Twitter can be a suitable source for investigating serendipity as a broad, everyday phenomenon, as opposed to focusing solely on the role of serendipity in scientific discovery or information seeking. To this end, we crawled a set of 30,000+ English-language tweets containing the word *serendipity* spanning a period of seven months (August 2011 through February 2012). Details of this data collection can be found in the first subsection below. Even in a relatively focused crawl such as this one, the presence of the word *serendipity* does not guarantee that the tweeter is describing a personal serendipitous experience. To better understand the different ways people use the concept, we performed a content analysis of a subset of our data set, which we detail in the second subsection.

Data collection

While there are many different ways of describing serendipitous occurrences and experiences in 140 characters or less, we focus on tweets containing the word *serendipity* in order to limit the need for manual filtering, as well as in order to understand how people, who actually know the term, describe their serendipitous experiences. We explored the use of the Twitter API to collect our data set, but this is

[1] http://blog.twitter.com/2012/03/twitter-turns-six.html

restricted to searching an index of between 6-9 days of the most recent tweets[2]. Related work suggests that serendipity is a relatively rare phenomenon (e.g., André et al., 2009), so to be able to estimate how often individuals experience serendipity, we needed to collect tweets over a longer time frame. We therefore used Topsy[3], a search engine for content posted on Twitter. Topsy's index contains tweets from as early as 2008, so we used Topsy to collect all tweets containing the word serendipity posted between August 1, 2011 and March 1, 2012. For this seven-month period Topsy contained 30,359 English-language tweets. Because Topsy is a text-based search engine, this automatically includes all 1,716 tweets that were tagged with the hashtag[4] *#serendipity*. We refer to this 30,000+ data set as TOPSY-ALL.

One problem with using Topsy for our tweet collection is that the Twitter API places a limit of 1% of the total amount of public tweets that can be accessed using the API. This also influences Topsy's indexing process. However, it is unlikely that the sample of Twitter that is indexed by Topsy, is biased towards or away from tweets about serendipity. We do not believe that these problems affect the conclusions we can draw from our data set collected using Topsy[5].

Coding tweets

To better understand how people tweet about serendipity, we performed a content analysis of a part of our data set and coded the tweets into different categories. To determine a list of appropriate coding categories, we took an open coding approach, where our coding categories emerge from the data (Lazar et al., 2010). Both authors developed their own coding categories based on a small set of 201 tweets published on February 1, 2012 taken from our original data set of +30,000 tweets, TOPSY-ALL. After calibration of our results, we merged our categories into a single coding scheme with five different categories:

- COMM tweets have a commercial intent, such as promoting jewelry, dresses or companies with the name Serendipity.
- LINK tweets contain links to Web content describing something related to the phenomenon serendipity.
- NAME tweets mention an object or location named Serendipity, such as movies, bars, restaurants, blogs, software, or bands.
- REFL tweets contain a general reflection, quote, or opinion about serendipity (but no clear description of a personal experience of a serendipitous occurrence).
- PERS tweets clearly describe a personal insight or experience of a serendipitous occurrence on the part of the tweeter[6].

These five categories are not mutually exclusive; a tweet about a personal serendipitous experience could be supplemented by a link describing the experience in more detail. Manually annotating all 30,000+ tweets in TOPSY-ALL was impractical, so we extracted two smaller data sets from on our original data that we coded for our content analysis phase. Both of these data sets contain only tweets that have been tagged with the hashtag *#serendipity*. This is a result of observations made during the development stage of our coding scheme: tweets with the hashtag *#serendipity* contain a greater number of PERS tweets than tweets simply containing the term serendipity, and PERS is the category we are most interested in.

The first and smallest of the two additional data sets, is TOPSY-150, which contains the 150 tweets published during the first 11 days of February 2012, tagged with *#serendipity*. In order to test our five-category coding scheme, both authors annotated all tweets in this TOPSY-150 data set. Since our five categories are not mutually exclusive, using annotation reliability measures such as Cohen's Kappa or Fleiss' Kappa is not appropriate (Lazar et al., 2010). We therefore calculated inter-annotator agreement

[2] See https://dev.twitter.com/docs for more information.
[3] http://www.topsy.com
[4] A Twitter hashtag is a short user-generated keyword prefixed with the hash symbol ('#') as a means of collating, sharing and following topics of interest in groups of users, who do not need to be connected through follower networks but take part in the same 'hashtag streams' (Bruns & Burgess, 2011). The same hashtag may be used for very different topics and events (as in our data set).
[5] An alternative to Topsy could be the TREC Microblog track's Twitter corpus (https://sites.google.com/site/trecmicroblogtrack/). However, we do not believe this to be more representative of Twitter than the 1% of tweets Topsy has access to.
[6] Only original tweets were coded as PERS; retweets were not considered to be personal serendipitous experiences of the person who retweeted the tweet in question. This is not 100% reliable as it is possible to hide the retweeted nature of a tweet.

(ITA) instead. Over all five categories combined, ITA is equal to 0.598. The ITA scores for the five categories are: LINK = 0.550; NAME = 0.633; COMM = 0.250; REFL = 0.414; PERS = 0.651.

Coding tweets is a subjective task that is often open to debate. For some categories the relatively low agreement is due to the lack of context associated with many tweets, making it difficult to unequivocally determine the tweeters' intent. However, for our main category of interest, PERS, we can consider the ITA as reflecting moderate agreement. After coding individually, we resolved any remaining differences through discussion to arrive at perfect agreement. When in doubt, we visited Twitter to inspect users' tweet histories, links, and relevant tweet conversations to disambiguate the tweet in question[7]. In our TOPSY-150 data set we ended up with the following category distribution: 1.3% of all tweets are classified as COMM, 27.3% as LINK, 39.3% as NAME, 18.0% as REFL, and 28.7% as PERS.

The next step in our annotation process was to analyze a bigger subset of TOPSY-ALL using our third data set, TOPSY-WINTER. This subset covers the three months of December 2011, January and February 2012 and contains 1073 tweets tagged with *#serendipity*. This means that the abovementioned data set TOPSY-150, which we used to test our coding scheme, is a subset of TOPSY-WINTER. In order to focus on our main category of interest (PERS) and to be able to code a larger set of tweets, we conducted a binary coding: tweets could be coded as PERS or as belonging to (at least) one of the other four categories (MISC). We did our best not to let established definitions of serendipity cloud our judgment; if a user clearly considered something to be serendipitous, then we marked it as PERS. After coding individually, we again resolved any remaining differences through discussion to arrive at perfect agreement. However, we were careful here not to overestimate: if a tweet was too ambiguous and/or contained too few contextual clues due to the 140-character limit of tweets, we assigned it to the MISC category[8]. This resulted in the following for TOPSY-WINTER: 160 tweets (or 14.9%) fall into the PERS category and 913 (85.1%) in the MISC category[9]. It is this TOPSY-WINTER data set that the majority of the work in the remainder of this paper is based on.

Experiencing Serendipity

In this section we present results aimed at answering what types of serendipity users experience and then communicate using Twitter (**RQ1**), and we examine three aspects related to **RQ1** in greater detail: (1) whether serendipitous experiences are leisure-related or work-related, (2) the different thresholds people have for calling something serendipity, and (3) the differences and similarities between the closely related concepts of serendipity and synchronicity.

Leisure vs. work

In order to understand the activities and contextual situations that accompany and influence people's serendipitous experiences, we performed a qualitative analysis of the 160 PERS tweets in TOPSY-WINTER, with a special focus on the distinction between leisure- and work-related activities. To increase our understanding of the context in the case of ambiguous tweets, we inspected users' tweet histories on Twitter to examine the tweets surrounding the ambiguous tweet in question.

A total of 141 PERS tweets (88.1%) were coded as leisure-related and 14 tweets (8.8%) as work-related. One tweet was coded as both: *I started typing 'An Engineer's Guide to Silicon Valley Startups' into Google and ended up watching this: http://t.co/YgNRA3XZ #serendipity* (id-J361). The work-related part about a startup guide is followed by a leisure-related link to a YouTube video found through Google's auto-fill suggestions: *"An Engineer's Guide to Cat Yodeling (with Cat Polka)".*

Many work-related tweets dealt with meeting people that might lead to new business opportunities, such as: *Just ran into @bruce_croxon, co-founder of lavalife. #serendipity* (id-J332). The last four tweets were too ambiguous to be classified into either category.

A closer inspection revealed a rich diversity in leisure-related activities connected to serendipitous experiences. As shown in Table 1, the identified activities cover all kinds of digital and physical spaces in everyday life containing affordances for encountering "information, objects, or people" (Rubin et al.,

[7] In TOPSY-150, we checked 40 of the 150 tweets (or 27%) in this detailed manner.

[8] For example: *Rest in Peace Sarah Marie... 2/20/09 #Serendipity* (id-F273). In subsequent examples, 'D' stands for December, 'J' for January, and 'F' for February in our internally assigned tweet IDs.

[9] The data sets used in this paper (TOPSY-150 and TOPSY-WINTER) are available at http://itlab.dbit.dk/~toine/?page_id=594. They contain Twitter IDs and Twitter user names for each tweet, our internally assigned IDs, and the result of our tweet annotation.

2011), including media, transportation, and shopping. The following examples may illustrate some of the rich diversity:

- Nature: *First sight of cherry blossoms this year! #serendipity http://t.co/UfYMKhOM* (id-F82). Link to an image of cherry blossoms in February.
- Search engine + music (YouTube): *Google found me a wrong Jung. And boy*, pleasant surprises and all. Almost a fan. Check http://t.co/Z6SFByo2 #Serendipity #guitarGeniouses* (id-D41). Link to a YouTube video with the South Korean guitarist Sungha Jung.
- People (chatting): *Checked the weather and found an old friend! I'd say that means "plenty of sunshine"! #Serendipity @CindyManzanero http://t.co/AjKFWgQO* (id-F280). Link to a smartphone screenshot of *'Baguio City Weekend Weather Chat'*.

Table 1
A sub-categorization of leisure activities connected to serendipitous experiences (N=160).

Category	Activities	Frequency[10]
Media (total)	-	77
(Media: Books)	book fair, book store, library, book reading/listening, book writing, word coincidences	(18)
(Media: Articles)	article reading, quote finding	(12)
(Media: Internet)	blogging, chatting, podcasting, Facebook, Twitter, Google alert, search engine, web surfing	(21)
(Media: Music)	live, radio, Spotify, YouTube	(21)
(Media: Movie)	movie, tv	(5)
People	face-to-face, email, phone, lecture	22
Transportation	car, commuting, travel, wayfinding, nature	30
Shopping	money-finding, shopping, gifts	20
Food + Household	cooking, eating (incl. restaurant), cleaning	15
Sports	watching, performing	7
Others	photographing, artwork, ngo activism, religion	5

The next subsections contain more examples of micro-serendipity tweets in our TOPSY-WINTER data set.

Serendipity thresholds

The qualitative analysis of the 160 PERS tweets in TOPSY-WINTER shows that there are clear differences in what could be called *serendipity thresholds*: when does a user find something unusual, unexpected, or surprising enough to consider it as serendipity? For example, this person links to an image of a pink balloon with the text 'Happy birthday princess': *Found this balloon on the side of the road. How fitting. #serendipity http://t.co/e2kxdlhS* (id-F228). The tweet mentioned in the previous section about the year's first sight of cherry blossoms (id-F82) is another example of how commonplace environmental occurrences can evoke serendipitous experiences. Being one of the few tweeters that posted more than one *#serendipity* tweet during the entire seven-month period covered by TOPSY-ALL (cf. section further below, 'Frequency of serendipity'), the same user also posted later on the same month: *First rainbow I've seen this year. #serendipity* (id-F173). These and similar observations in our data set indicate that different people have different thresholds for what they consider serendipitous; plain novelty or pleasant diversion may be enough. This finding echoes what Makri & Blandford (2012) emphasize about serendipity being a subjective phenomenon.

[10] These categories are non-exclusive, i.e., some tweets cover two or more activities at the same time.

Serendipity vs. synchronicity

When Horace Walpole coined the word 'serendipity' in 1754 as "making discoveries, by accident & sagacity, of things which they were not in quest of" he stressed that "no discovery of a thing, you are looking for, comes under this description" (Van Andel, 1994 , p. 633).

In TOPSY-WINTER, 18.8% of our tweeters use the term serendipity when they, in unexpected ways, discover something they were already looking for, or in other ways were preoccupied with, such as: *Walked into a mall I don't know that well and choose the entrance right by the store I'm here for. #Serendipity* (id-D212). These kinds of experiences related to a preoccupied mind can also be seen in the following typical example: *I joke to colleagues about feeling very Flashdance today & Spotify sends me: What A Feeling - Irene Cara #serendipity http://t.co/PfrzG6UG* (id-J222).

It is clear from both our analysis of micro-serendipity and from related work (Makri & Blandford, 2012) that the notion of serendipity has broadened to encompass situations like the above. As noted by Makri & Blandford (2012), this kind of experience is also called *synchronicity*, which Wikipedia defines[11] as "the experience of two or more events that are apparently causally unrelated or unlikely to occur together by chance, yet are experienced as occurring together in a meaningful manner." Many of the serendipitous experiences involving music in the data set fall in this 'synchronicity' category, like in the *Flashdance* example above.

In the framework of the current research project we define synchronicity as a match between a perceived accidental occurrence in a person's environment and a foreground activity, problem, or interest preoccupying that person[12]. In concordance with traditional definitions (cf. Merton & Barber, 2004; Van Andel,1994), traditional serendipity may correspondingly be defined as a match between a perceived accidental occurrence in a person's environment and a background interest that can be triggered in that person. We elaborate on this difference further below in the sections 'Describing serendipity' and 'Discussion & Conclusions'.

Frequency of serendipity

Traditional studies of serendipity (e.g., André et al., 2009) have suggested that it is a relatively rare and anomalous phenomenon. While many people may have experienced serendipity, the same person may experience serendipity on a relatively infrequent basis. In this section we use our data sets to provide a realistic estimate of the frequency of serendipitous experiences shared on Twitter (**RQ2**).

One way of looking at this is from the perspective of Twitter as a whole: can we come up with a cautious estimate of what proportion of tweets describes serendipitous occurrences? According to Twitter, they processed 340 million tweets per day in March 2012 vs. 200 million in June 2011[13]. A linear extrapolation[14] of these numbers to the three-month period covering TOPSY-WINTER would mean an average of 305 million tweets per day in that period. English makes up approximately 51.5% of all tweets[15] (Hong et al., 2011), which gives us about 157 million English language tweets per day.

In TOPSY-WINTER 160 out of 1073 tweets fall in the PERS category (14.9%). This amounts to an average of 1.8 PERS tweets per day that have been tagged *#serendipity*, as indexed by Topsy. Even if Topsy has access to just 1% of all Twitter's tweets, this means that tweets describing serendipitous occurrences make up a vanishingly small proportion of the grand total of 157 million tweets. However, this is likely to be an underestimate, as we may assume that a considerable number of people tweet about their serendipitous experiences without tagging them *#serendipity*.

A better way of examining the frequency of serendipity may be within a personal context: how common is it for an individual to tweet about a personal serendipitous experience? If we look within our TOPSY-WINTER data set, we find that 146 different users account for 160 PERS tweets about a personal serendipitous experience using the hashtag *#serendipity*. Of these users, three sent out a pair of two identical tweets on different dates each. One other user retweeted the same tweet on eight different

[11] http://en.wikipedia.org/wiki/Synchronicity
[12] This variant kind of serendipity has also been called *pseudo-serendipity* (e.g., Van Andel, 1994).
[13] http://blog.twitter.com/2011/06/200-million-tweets-per-day.html
[14] This is likely to be a conservative estimate given Twitter's super-linear growth in the past as measured by the number of tweets per day.
[15] This is assuming all languages grow at an equal rate, which is an oversimplified assumption.

dates, accounting for a total of 160 tweets. This means that in our three-month TOPSY-WINTER data set each user has only tweeted about a single unique serendipitous experience[16].

This does not allow for any general conclusions to be drawn about the personal frequency of serendipity shared on Twitter. This could suggest that serendipity is a rare phenomenon—or perhaps a rarely communicated phenomenon—and that a larger data set is likely required to determine its frequency more reliably.

Describing serendipity

In this section we examine how tweeters describe their personal serendipitous experiences using Twitter: is there such a thing as a vocabulary for serendipity (**RQ3**)? We examine word usage, parts-of-speech, and Twitter hashtags associated with PERS tweets in our TOPSY-WINTER data set.

Word usage

In the first subsection below we make a quantitative analysis using a log-likelihood statistic on the data set. This is supplemented in the second subsection with a qualitative analysis of the data set.

Quantitative analysis. If we can determine the vocabulary commonly used to describe personal serendipitous experiences, it could help us identify other serendipity-related tweets in addition to the ones containing the word *serendipity* or the hashtag *#serendipity*. The presence or absence of such 'serendipitous' words could, for instance, be used as features for constructing a classifier that flags possibly serendipity-related tweets. In our situation we wish to determine whether certain terms are more characteristic for PERS tweets than for MISC tweets. This is similar, albeit on a smaller scale, to determining whether there is difference in word distributions between two text corpora. A robust measure for determining the surprise of a word's usage between two corpora is log-likelihood as proposed by Dunning (1993). Table 2 shows the 30 most characteristic terms in the TOPSY-WINTER data set for both the PERS and MISC tweets, ordered by log-likelihood.

Table 2
Top 30 most representative terms for the PERS and MISC tweets in the TOPSY-WINTER *data set as indicated by their log-likelihood score.*

Tweet category	Top 30 most representative terms
PERS	*just, found, road, met, book, walked, spring, ran, noticed, chinese, charity, car, note, today, song, store, simultaneously, shop, radio, pleasant, oooo, omg, immediately, bumped, picture, named, heard, flowers, flight, heard*
MISC	*watching, serendipity, hot, chocolate, movie, frozen, love, excited, welcome, cleversense, others, checked, create, beautiful, movies, discovery, spur, kate, chance, network, john, christmas, watch, search, panel, fave, york, sundae, heart, museums*

A few of the PERS terms can be expected to be representative of tweets describing serendipitous experience, such as *just*, *found*, *noticed*, *bumped*, *simultaneously*, *immediately*, and *omg* (i.e., "oh my god"). The TOPSY-WINTER data set contains a lot of tweets mentioning the movie *Serendipity*, the *Serendipity3* restaurant in New York, and blog posts about how innovation can lead to discovery and serendipity. This is reflected in the MISC terms, such as *movie(s)*, *kate/john*[17], *frozen*, *chocolate*, and *discovery*. Paradoxically, the word *serendipity* is much more indicative for MISC tweets than for PERS tweets due to its use in product names, which emphasizes the importance of finding other ways of detecting how people signal serendipitous experiences. In general, however, we believe TOPSY-WINTER to be too small for us to be able to discover a true vocabulary of serendipity unsupervised.

[16] We extended the same analysis to our entire TOPSY-ALL data set, where a handful of these 146 users tweeted about a serendipitous occurrence more than once. The majority, however, did so only once in the seven-month period.
[17] Kate Beckinsale and John Cusack are the lead actors in the movie *Serendipity*.

Rubin et al. (2011) compiled a list of 43 queries meant to identify blog posts about serendipitous occurrences, suggesting there is a certain terminology associated with serendipity. We received the full set of serendipity queries[18] from Rubin et al. and treat them as another set of tweets to see if there is an overlap between the query vocabulary and the PERS and MISC tweet sets from TOPSY-WINTER. We calculate Kullback-Leiber-divergence (or KL-divergence) (Manning & Schütze, 1999) between the three different word distributions. KL-divergence is a measure of the similarity between two distributions; in our Twitter scenario, the lower the KL-divergence is, the more alike two text collections are in their word usage.

The KL-divergence between the PERS and MISC tweets is equal to 13.33. However, the KL-divergence between Rubin's queries and the PERS tweets is equal to 20.31, whereas the KL-divergence between Rubin's queries and the MISC tweets is equal to 20.25. The PERS and MISC tweets are thus much more alike in word usage than Rubin's manually queries are to either of the two sets of tweets. In addition, there is virtually no difference between the KL-divergences of Rubin's queries to the two sets (20.31 vs. 20.25). Rubin's queries can therefore not be used to reliably distinguish between the two tweet categories and thus identify serendipitous tweets. This suggests that Twitter vocabulary is indeed different from the vocabulary used on blogs for describing serendipitous occurrences.

Qualitative analysis. Given the lack of clear results using log-likelihood, we took a closer qualitative look at the actual words signaling serendipity in the 160 PERS tweets in TOPSY-WINTER. Manually marking up these terms revealed a pattern of four key elements as shown in Table 3: *preoccupation*, *unexpectedness*, *insight* and *value*. The latter three elements were taken from a study by Makri & Blandford (2012), whereas *preoccupation* is a fourth key element we identified in our analysis. It is also discussed by Makri & Blandford (op.cit.), but not recognized as a key element.

Table 3
Examples of manually identified terms signaling serendipity, grouped by key serendipity elements.

Preoccupation	*continues* (F123), *I have been listening to … almost exclusively for a week* (D68), *I was just* (J288), *not the first time* (F381), *read prior* (J65), *the same […]* (F41)
Unexpectedness	*actually* (J351), *and lo* (J90), *and look* (F131), *and then today* (D68), *bumped into* (D137, J368), *came across* (J10), *came on* (F185), *catch* (D13), *digging into* (D35), *dropped into* (F102), *find* (J282, F28/118), *found* (D40/41/265 + 8 more), *got it* (J375), *happen to have* (D297), *happens to be* (J287), *happened to get* (J286), *if not … we would've missed* (J357), *just* (D46/137 + 17 more), *lucky* (F165), *met* (J351, F360), *omg* (F343), *popped up* (D46), *ran into* (J332, F110/240), *stumbled upon* (J203), *surprises* (D41), *there it was* (D21), *unexpected* (F301), *walked into* (J152/267), *while looking for something else* (J10, F358)
Insight	*and I see* (D185), *discover* (F178), *discovered* (J424), *found out* (J287), *haven't seen it either* (F149), *hunch* (F47), *look at* (J286), *noticed* (J344, F47/144), *realized simultaneously* (J414), *immediately thought* (F97)
Value	*amazing* (J152), *appropriate* (F332), *awesome* (D156), *been looking for since* (F310), *cool* (D165, J417), *excellent* (F48), *'!'* [exclamation mark] (F82 + many more), *favorite* (F85), *fitting* (F228), *free* (D276), *good* (D21, J147), *great* (J368/420), *joy* (D13), *just in time* (F291), *love the web* (D140), *love it* (J110, F364), *love this* (F93), *pleasant* (D41), *score!* (D156), *smile* (J147), *:-)* [smiley] (F76), *so perfect* (F102), *tasty* (D148), *timing perfectly* (J394), *what a lovely* […] (D109), *whoa!* (F120)

Table 3 shows that close inspection of the TOPSY-WINTER PERS tweets confirms the presence of all three key serendipity elements of *unexpectedness*, *insight* and *value* identified by Makri & Blandford, although not always explicitly so at the same time.

Sometimes just the element of unexpectedness could be identified with the elements of insight and value only implicitly present. This is illustrated by the following example: *Cool! Just when I was*

[18] Example queries include *"found some * * by accident"* and *"found * * by accident"* (Rubin et al., 2011).

wondering what to get the wife for Valentine's Day! #Serendipity http://t.co/5QibILEW (id-J417, link to image of dish drying rack). The element of *unexpectedness* is denoted by the word 'just' and the exclamation mark. *Insight* is shown by the link pointing to a humorous idea of a Valentine's gift. The *value* is signaled by the word 'Cool', both exclamation marks, and the simple fact that the tweet has been shared on Twitter.

In another example: *Realized simultaneously that I have no wrapping paper & the Chinese place left menu on my door. #serendipity #SorryDad* (id-J414), *unexpectedness* is signaled by stating that, surprisingly, there is 'no wrapping paper' and that a solution to this problem is immediately and equally surprisingly found. The *insight* lies explicitly in the wording 'realized', whereas the implicit *value* is that the gift ended up being wrapped.

Our close inspection of the tweets suggested a fourth key element of serendipity: the user's possible *preoccupation* (i.e., foreground problem/interest). This is in line with the results presented in the previous section ('Serendipity vs. synchronicity') about synchronicity. The following example illustrates how all four serendipity elements come together in some of the tweets: *Found this balloon on the side of the road. How fitting. #serendipity* (id-F228). As earlier mentioned, this tweet linked to an image of a pink balloon with the text 'Happy birthday princess'. The *preoccupation* here is that the tweeter most likely knows a girl having birthday within a short time range. *Unexpectedness* is shown by the terms 'found on the side of the road'. Both *insight* and *value* are expressed with the expression 'how fitting'.

We manually compared the terms in Table 3 with the final set of 16 queries that Rubin et al. (2011) used for retrieving blog posts about serendipitous experiences. Their queries were permutations of the phrases *"looking/searching for [...] but found/discovered", "stumbled across/found [...] by chance/accident [...] looking for", "wasn't looking [...] but/when [...] found", "found [...] while looking"* and *"accidently found"*.

As shown in the previous subsection, there are differences with the terms used in our data set. A manual comparison showed that, while there is some overlap (e.g., 'found' and 'stumbled'), Rubin et al. do not cover several of the variations in word usage present in our data set, e.g. 'bumped into', 'came across', 'happened to get', and 'stumbled upon'. Twitter data revealing users' actual word usage could thus suggest terms for other serendipity studies.

Parts-of-speech

Another aspect of the vocabulary of serendipity is which parts-of-speech (POS)—the lexical category of a word—are most often used to describe serendipitous occurrences. Whereas word usage tends to be much more varied, there is only a limited set of POS tags that can be assigned to a word, which could lead to clearer patterns in the description of PERS tweets. For POS tagging we used the MBSP toolkit (Daelemans & van den Bosch, 2005). We extracted the POS tags from our TOPSY-WINTER data set, for both the PERS and the MISC tweets, and filtered out stop words, symbols, and cardinals. This resulted in the following top 9 most frequent POS categories, shown in Table 4.

We find three interesting deviations from the normal distribution of POS categories. Nouns are about 10% more likely to occur in MISC tweets than in PERS tweets, which match our earlier observations that many tweets marked with *#serendipity* contain mentions of movies, bars, restaurants, and companies called *Serendipity*. In contrast, past tense verbs are twice as prominent in PERS tweets compared to MISC tweets. A likely explanation for this is that tweeting about serendipitous occurrences involves describing past events, necessitating the use of past tense verbs. Tweeting about watching, consuming or experiencing 'resources' called *Serendipity* are more often described in present tense as the event takes place, as evident from the 44% higher occurrence of present tense verbs. While none of these POS categories individually are enough to identify PERS tweets, these combined distribution patterns do show promise as features in a classifier for tweets about serendipitous occurrences.

Table 4
Distribution of the most frequent, non-stopword POS categories for PERS and MISC tweets.

POS category	Relative frequency in PERS	Relative frequency in MISC
Noun	34.80%	38.14%
Adjective	7.90%	8.85%
Adverb	7.35%	4.70%
Determiner	6.88%	4.68%
Pronoun	5.85%	6.13%
Verb, past tense	5.61%	2.83%
Verb, present tense	2.21%	3.19%
Verb, stem	2.17%	3.62%

Hashtagging serendipity

A third way people can signal serendipitous occurrences is by using hashtags on Twitter. Our analysis of our TOPSY-WINTER data set shows that *#serendipity* is used by many users to describe serendipitous occurrences, but can we identify other hashtags that serve a similar purpose? A first step would be to identify the hashtags most frequently co-occurring with *#serendipity* for the PERS tweets in TOPSY-WINTER. Due to the relatively small size of TOPSY-WINTER and the fact that hashtags in general occur less often than normal words in tweets, the results of such an unsupervised identification are disappointing. The most commonly co-occurring hashtags typically represent locations and events, such as *#nyc*, *#superbowl*, *#saints*, and *#weezercruise*.

We can try to address the problem of data set size by looking at TOPSY-ALL: using log-likelihood we have determined the most characteristic hashtags that co-occur with *#serendipity*. An unsupervised approach does not seem to identify the most promising hashtags here either: *#serendipity* is often used to signal new content to a particular crowd or to describe the names of products or places, so the top of the list is dominated by such occurrences due to 'spam'. However, a manual inspection of the list does reveal potential other serendipity-signaling hashtags, such as *#synchronicity*, *#serendipitous*, *#chance*, *#insight*, *#wtf*, *#randomness*, *#accident*, *#lucky*, and *#surprise*. These could be used in future work to collect additional tweets describing serendipitous occurrences. A fruitful method for future harvesting such tweets could for instance be scoring tweets by the number of these 'signal' hashtags they are tagged with and then select the highest-scoring tweets.

Discussion & Conclusions

In this paper, we have presented our work on *micro-serendipity*: investigating everyday contexts, conditions, and attributes of serendipity as shared on Twitter. One aspect of our investigation focused on how people share their serendipitous experiences on Twitter (**RQ1**). We found that users have very different thresholds for considering something serendipitous. We observe tweeters using the term serendipity for everyday occurrences ranging from pleasant diversions and distractions to wholly unexpected and unusual events. We propose generalizing this to a *serendipity continuum* to cover the entire spectrum of different degrees of surprise, from unplanned everyday incidents to unanticipated eureka moments in science. This is in line with Makri & Blandford (2012), who argue against viewing serendipity as a purely discrete concept.

While we cannot say anything conclusively about how often people experience serendipity in real life, we find that sharing such experiences via Twitter is relatively rare (**RQ2**). One explanation could be that serendipity is a rare phenomenon in general. However, people seem to have different thresholds for considering events serendipitous, which could mean that for some serendipity might be so common an everyday phenomenon that they do not always reflect upon it or find it worthwhile to share with others. Future work dealing on **RQ2** would require annotating a larger collection of tweets taken at different points in time, possibly combined with an analysis of some individual hashtag streams (Bruns & Burgess, 2011) to determine which of these two explanations is most likely.

We do not have a conclusive answer to the question whether there is such a thing as a characteristic vocabulary of serendipity on Twitter (**RQ3**). On the relatively small scale of TOPSY-WINTER, the inherent variation of natural language dominates any patterns that might occur in the data. An analysis of the POS categories in PERS and MISC tweets revealed some interesting differences, whereas the analysis of hashtag usage was again inconclusive on its own due to the size of TOPSY-WINTER. However, while none of these three aspects of describing serendipity are characteristic enough on their own to detect tweets about serendipitous occurrences, we believe that *combining* these different serendipity-signaling features could be a promising approach to automatically detecting tweets about serendipitous occurrences. Tweets surrounding a suspected serendipitous tweet could also serve as extra context features here. Future work could involve training such a classifier for identifying such occurrences automatically. Techniques such as active learning (Manning & Schütze, 1999), where a small, annotated data set can be expanded in a continuous feedback loop of extracting features and detecting new patterns, may be useful here.

Our qualitative analyses revealed a pattern behind the diversified and punch-line word usage of tweets, confirming the presence of *unexpectedness*, *insight*, and *value* as identified by Makri & Blandford (2012). All three elements were present with different degrees of explicitation. Our study suggests that the person's possible *preoccupation* could be included as a fourth serendipity element, thereby covering the aspect of *synchronicity* also present in the study by Makri & Blandford (op.cit.), but not included in their key elements. Just like the researchers interviewed by Makri & Blandford, the tweeters in our data set expressed a broad view on types of serendipity. In order to cover this range of serendipity types, we propose referring to the 'traditional' serendipity concept as *background serendipity*, because it is characterized by unexpectedly finding something meaningful related to a *background* interest, thereby *changing* that person's focus and direction. The other frequent type of serendipity, not only experienced in everyday life, but also in science (op.cit.), is what we correspondingly refer to as *foreground serendipity* (i.e., synchronicity), as it is characterized by unexpectedly finding something meaningful related to a *foreground* interest and preoccupation, thus *confirming* the person's focus and direction.

Both types of serendipity deal with people experiencing *meaningful coincidences*; in other words, with people considering an occurrence as *both* incidental *and* meaningful. This very *consideration* by the person matches what Makri & Blandford (op.cit.) call *insight*, and Horace Walpole called *sagacity* in 1754 (Merton & Barber, 2004; Van Andel, 1994). An occurrence must thus be *considered* as both incidental and meaningful in order for a person to denote it as serendipitous. Therefore, the two most important elements constituting a serendipitous experience seem to be *unexpectedness* and *value*—i.e., the *meaningful coincidence*—as *preoccupation* is not always current and some degree of *insight* must always be present in order to consider an occurrence as both unexpected/incidental and valuable/meaningful. Our study suggests that the respective thresholds for considering something as incidental and meaningful are highly subjective, explaining the range of the aforementioned *serendipity continuum*.

Summing up, Table 5 gives an overview of the key elements in *background serendipity* and *foreground serendipity* as discussed above.

Table 5
Key elements in background serendipity and foreground serendipity.

	Background serendipity	Foreground serendipity
Preoccupation	–	+
Unexpectedness	+	+
Value[19]	+	+

In future work we will look more closely at how tweeters describe matches between environmental factors and their foreground/background interests. By casting serendipity as a correspondence between environmental and personal factors, we will extend our conceptual framework to include affordance theory as suggested by Björneborn (2010). In this framework, serendipity can be

[19] The elements of 'unexpectedness' and 'value' Include 'insight' (Makri & Blandford, 2012) as discussed in the text.

seen as an affordance, i.e., as a three-way relationship between an environment, a human being, and a potential activity (cf. Dourish, 2004, p.118).

The work by Rubin et al. (2011) on analyzing blog posts that describe serendipitous occurrences served as an inspiration to the work described in this paper. While we have identified several similarities between the two approaches, there are also some key differences. By fine-tuning their queries, Rubin et al. aim for a precision-oriented approach to discovering anecdotal evidence of serendipity, based on a narrow, preset definition of the concept. In contrast, we cast a wide net and aim to identify more of the possible variations in the use of serendipity on Twitter, resulting in a data set spanning both information-related serendipity as well as everyday occurrences. While our data is more ambiguous due to the shorter length of tweets as compared to blog posts, and contains fewer quality descriptions of serendipitous occurrences according to the traditional definition, it does showcase the diversity in word usage and use of the concept to a much higher degree. Twitter data revealing users' actual word usage could thus suggest terms for other serendipity studies.

Even if tweeters are afforded much less space for describing their inner thoughts than, for instance, bloggers, tweets do allow for more unfiltered (Bruns & Burgess, 2011), spontaneous, and near-instantaneous inspection, approaching real-time 'streams of consciousness' (as opposed to the retrospective nature of earlier serendipity research). If one exercises caution when filtering Twitter data, combined with close inspection as shown in our data analysis, we believe Twitter to be a promising resource for research into how people experience everyday life including micro-serendipity.

References

André, P., schraefel, m. c., Teevan, J., & Dumais, S. T. (2009). Discovery is Never by Chance: Designing for (Un)Serendipity. In *Proceedings of the 7th ACM Conference on Creativity and Cognition* (pp. 305–314). Berkeley, CA: ACM Press.

Björneborn, L. (2010). Design Dimensions Enabling Divergent Behaviour across Physical, Digital, and Social Library Interfaces. In *Proceedings of PERSUASIVE 2010* (pp. 143–149). Springer.

Bruns, A., & Burgess, J. (2011). The Use of Twitter Hashtags in the Formation of Ad Hoc Publics. In *Proceedings of the 6th European Consortium for Political Research General Conference*.

Daelemans, W., & van den Bosch, A. (2005). *Memory-Based Language Processing*. Cambridge University Press.

De Rond, M. & Morley, I. (Eds.) (2010). *Serendipity: Fortune and the Prepared Mind*. Cambridge University Press.

Dourish, P. (2004). *Where the Action is: The Foundations of Embodied Interaction*. MIT Press.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), 61–74.

Erdelez, S. (2004). Investigation of Information Encountering in the Controlled Research Environment. *Information Processing & Management*, 40(6), 1013–1025.

Hong, L., Convertino, G., & Chi, E. H. (2011). Language Matters in Twitter: A Large Scale Study. In *Proceedings of ICWSM '11*.

Lazar, J., Feng, J. H., & Hochheiser, H. (2010). *Research Methods in Human-Computer Interaction*. John Wiley & Sons.

Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Makri, S., & Blandford, A. (2012). Coming across Information Serendipitously: Part 2 – A Classification Framework. *Journal of Documentation*, 68(5), 706–724.

McCay-Peet, L., & Toms, E. G. (2011). Measuring the Dimensions of Serendipity in Digital Environments. *Information Research*, 16(3), Paper 483. Retrieved from http://InformationR.net/ir/16-3/paper483.html

Merton, R. K., & Barber, E. (2004). *The Travels and Adventures of Serendipity: A Study in Historical Semantics and the Sociology of Science*. Princeton University Press.

Piao, S., & Whittle, J. (2011). A Feasibility Study on Extracting Twitter Users' Interests using NLP Tools for Serendipitous Connections. In *IEEE Third International Conference on Social Computing (SocialCom)* (pp. 910–915).

Rubin, V. L., Burkell, J., & Quan-Haase, A. (2011). Facets of Serendipity in Everyday Chance Encounters: A Grounded Theory Approach to Blog Analysis. *Information Research*, 16(3), Paper 488. Retrieved from http://InformationR.net/ir/16-3/paper488.html

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter Events. *Journal of the American Society for Information Science and Technology*, 62(2), 406–418.

Van Andel, P. (1994). Anatomy of the Unsought Finding: Serendipity: Origin, History, Domains, Traditions, Appearances, Patterns and Programmability. *British Journal for the Philosophy of Science*, 45(2), 631–648.

Zhang, Y., Séaghdha, D., Quercia, D., & Jambor, T. (2012). Auralist: Introducing Serendipity into Music Recommendation. In *Proceedings of the 5th ACM Conference on Web Search and Data Mining, WSDM-12*.