Using social media to find English lexical blends[1]

Paul Cook

Keywords: *lexical blends*, *neologisms*, *computational lexicography*, *social media*, *Twitter*.

Abstract

We present a method for identifying English lexical blends — words such as *complisult* (*compliment* + *insult*) and *globesity* (*global* + *obesity*) — from social media, specifically Twitter. Our method is based on observations about words and phrases that are commonly used to introduce new words and corpus patterns that are often used to describe the meaning of lexical blends, and leverages the massive volume of data that is readily-available for analysis through Twitter. We run our method for 5 weeks and identify 976 candidate lexical blends; analysis of a sample of these candidates indicates that approximately 57% are blends. We further discuss a small number of blends identified by our method that are in regular usage on Twitter but which are not recorded in any of a number of dictionaries searched.

1. Lexical blends

Lexical blends are words such as *glamping* (*glamorous* + *camping*), *sexting* (*sex* + *texting*), and *sleepiphany* (*sleep* + *epiphany*) that are formed by combining a prefix of one source word with a suffix of another. Blends exhibit some variation with examples being observed in which neither source word is, one source word is, and both source words are entirely present in the blend, as in *glamping*, *sexting*, and *sleepiphany*, respectively. Blends can also be formed from proper nouns, for example, *Movember* (*moustache* + *November*), and although blends typically combine a prefix of one word with a suffix of another, blends of three words, for example, *turducken* (*turkey* + *duck* + *chicken*), and blends in which one word (or part of a word) is inserted into another, for example, *entreporneur* (*entrepreneur* + *porn*), are also observed.

While blends are cute, creative formations, they are also a common type of new word. For example, in studies drawing on very different data sources, Algeo (1991) finds roughly 5% of the new words he analyzes to be blends, while blends accounted for roughly 43% of the neologisms examined by Cook and Stevenson (2010). There has also been considerable recent interest in the linguistic study of blends (e.g., Renner et al., In preparation).

The users of online social media produce a tremendous amount of text each day, much of which is readily available for lexicographical analysis. Easy access to such very large corpora is quite new, and has opened new possibilities for lexicographical inquiry. In particular, corpus patterns that are very rare in conventional-size corpora turn out to have many occurrences in the very large corpora of social media.

The goal of this study is to consider whether the massive amounts of text available through Twitter — a popular micro-blogging service, discussed in Section 2 — and simple computational processing, can be used to help to identify English lexical blends; we are primarily interested in finding blends that are in usage, but that are not recorded in dictionaries. The methods we propose could be applied to construct a large dataset of lexical blends, which would be a great asset to the linguistic study of blends.

After discussing previous observations about the contexts in which neologisms and lexical blends are often used in Section 3, we present our method for identifying blends in Section 4; this method is based on searching corpora for very rare patterns that often indicate that a given word is a blend. The output of this process is a list of candidate lexical blends; in Section 5 we analyze this output and find that approximately 57% of the candidates are indeed

blends. We further discuss a small number of the blends identified by our system which are not yet recorded in any of a number of dictionaries searched. We briefly discuss two alternative approaches to identifying blends in Section 6, and finally offer some concluding remarks in Section 7.

2. Twitter

Twitter is a micro-blogging service that allows users to post short messages, typically publicly available, of maximum length 140 characters. Tweets — messages posted to Twitter — tend to be of a rather informal register; Twitter is therefore an excellent source of data to search for new blends because it is expected to contain many expressions that would be unlikely to occur in more formal registers. Furthermore, a very large amount of data is available, with Twitter reporting that as of June 2011 roughly 200 million tweets were being sent each day (Twitter, 2011). Crucial to this study, Twitter supports research (and other applications) by allowing easy access to its data.

3. Related work

Barrett (2006) discusses an approach to finding lexical items that are new or unrecorded in dictionaries by searching for phrases that often indicate that a word is outside of an author's vocabulary. For example, in the following tweets *another word for*, *referred to as*, and *new term* are indications that the boldface terms are unfamiliar to the authors of these tweets. (In examples of tweets usernames, names, and URLs have been replaced with <USER>, <NAME>, and <URL>, respectively.)

(1) @<USER> **swag** is another word for style. Swagging B)

(2) Heard great new term on Today programme. '**Sodcasting**' playing your music in public so loud that others can hear.

(3) Urban Barns Foods has a unique style of agriculture referred to as **Cubic Farming**. <URL>

Indeed, Barrett (2006) used this process (although not applied to Twitter data) in the writing of a dictionary of lexical items that were otherwise undocumented or under-documented. This approach has the potential to identify lexical items that are very infrequent, which we expect many lexical blends to be. Furthermore, because of the large amount of data available on Twitter, we can search for, and expect to find, many usages of phrases that indicate that a lexical item is new to an author, even if those phrases are rather infrequent in conventionally-sized corpora. However, this approach is not able to specifically identify lexical blends.

Cook and Stevenson (2010) present a preliminary statistical approach to automatically distinguishing lexical blends from other types of words. Specifically, for a given target word (blend or otherwise) they identify all pairs of words *w1*, *w2* such that — based only on orthography — if the target were a blend its source words could be *w1* and *w2*. They then assign a numerical score to each of these candidate pairs of source words; this score is formulated to be high for pairs that are plausible source words according to a number of previously-observed linguistic properties of blends, and low otherwise. Cook and Stevenson show that blends tend to have a higher scoring candidate pair than non-blends; however,

although of theoretical interest, their method is quite computationally expensive, and so is not well-suited to identifying blends in the large amounts of text available through social media.

Cook (2010) notes that the meaning of a blend is often indicated by its source words occurring nearby in text. For example, the meaning of *subtweet* (*subliminal + tweet*) is explained in the following tweet.

(4) @<USER> it basically means subliminal tweet .. so its subtweet for short

In the following section we present a method for identifying lexical blends that draws on this observation, and Barrett's (2006) observations about phrases that indicate that a word is new.

4. Methods

Our approach to finding lexical blends involves two steps: we first identify tweets containing words or phrases that indicate that a word is new to an author, and we then identify candidate blends amongst those tweets.

We begin with a set of 29 regular expressions matching words and phrases that often indicate that a word is new to an author. This list was formed based on examples of such expressions from Barrett (2006) and by examining entries in the Double-Tongued Dictionary (also edited by Barrett, http://www.doubletongued.org/). Examples from this list, along with examples of strings matching these regular expressions, are given in Table 1.

Table 1. Examples of regular expressions and matching strings.

Regular expression	Examples of matching strings
coined the (term|word)	*coined the term, coined the word*
jargon for	*jargon for*
known in \w+ (terms|speak)	*known in technical terms, known in computer speak*
known to \w+ as	*known to scientists as, known to geeks as*
new word	*new word*
slang (expression|phrase) for	*slang expression for, slang phrase for*

The Twitter Streaming API (application programming interface, https://dev.twitter.com/) allows us to easily obtain tweets containing specific keywords as they are tweeted. We form queries of keywords corresponding to the 29 regular expressions that often indicate that words are new. For example, for the regular expression 'coined the (term|word)' we form two queries: *coined the term* and *coined the word*. For the 29 regular expressions this results in a total of 58 queries which we then obtain tweets for using the API. The API provides us with tweets containing all of the keywords in a query, but does not guarantee the order of the keywords in those tweets. We therefore use a simple Python program to post-process the resulting tweets to keep only those matching one of the original 29 regular expressions. (See Bird et al. (2009) for an introduction to regular expressions in the context of natural language processing using Python.) We ignore case in all the processing describe in this section.

From the remaining tweets we then use another simple Python program to identify those which potentially contain a lexical blend and its source words. Specifically, we search for any tweet containing a candidate blend *b*, and 2 other words *w1* and *w2* , such that *b* could

potentially be a blend of *w1* and *w2*, that is, *b* can be written as a prefix of *w1* followed by a suffix of *w2* (where here we use the terms *prefix* and *suffix* to refer to the beginning or ending of a string, not morphological affixes). For example, the following tweet is identified as having a candidate blend *carbage* with candidate source words *car* and *garbage*.

(5) Coined new term today for all the garbage and random stuff in my car: "carbage"

To reduce the number of non-blends that are identified as candidate blends by our system, we further introduce a number of simple heuristics. These heuristics will tend to exclude types that are unlikely to be lexical blends, but will of course in some cases cause us to fail to identify some blends. We discuss some of these heuristics, and their shortcomings, below.

We require that a candidate blend consist of only alphabetic characters, and not be included in a large wordlist. (Specifically, we use the wordlist from GNU Aspell 0.60.6.1.) Because g-clipping (e.g., *talkin* for *talking*) is so common on Twitter, we expand the wordlist to include such forms. This wordlist restriction prevents our method from identifying many (in-vocabulary) non-blends that happen to appear in contexts that could be explaining the meaning of a blend. However, this restriction does prevent us from identifying blends that are homonyms of words in the wordlist. For example, *chugger* is a blend of *charity* and *mugger*, but this lemma also has a non-blend usage related to fishing, which is included in some dictionaries (e.g., "chugger, n.". OED Online. December 2011. Oxford University Press. 29 February 2012 http://www.oed.com/view/Entry/264315?redirectedFrom=chugger). (*Chugger* happens to not be included in the wordlist we use, so this particular case is not a problem for us; this example is only intended to illustrate the kind of problem that could be encountered.) We further require that a candidate blend be at least six characters in length because shorter words are more likely to have a possible interpretation as a blend than longer words. This length restriction prevents short non-blends from being incorrectly identified as blends, but this of course comes at the cost of not being able to identify shorter blends, for example, *spork* (*spoon + fork*).

We further restrict the blends identified according to a number of heuristics related to their candidate source words. Following Cook and Stevenson (2010), we require that both candidate source words contribute at least two characters to a candidate blend. Again this heuristic prevents an overly-large number of non-blends from being identified, but comes at the cost of missing any blend in which a source word only contributes a single letter, for example, *vortal* (*vertical + portal*). We also require that the stem of neither candidate source word — as obtained from the algorithm of Porter (1980) — correspond to the stem of the candidate blend or the blend itself, and furthermore that the material that the second candidate source word contributes to a candidate blend not correspond to an inflectional suffix. We further require that the first candidate source word not be a prefix of the second candidate source word, and that the second candidate source word not be a suffix of the first candidate source word.

In Twitter, words are commonly lengthened, often for emphasis, as in the following example.

(6) aaaaaaaaaaaah I looove this...<NAME> touching <NAME>'s faceeeee

Such lengthened forms are unlikely to be lexical blends. To avoid identifying such forms as blends, we use a heuristic inspired by Kaufmann and Kalita (2010), and require that all candidate blends not have a sequence of more than three repeated characters.

The focus of the present study is English lexical blends. The tweets matched by our

regular expressions will tend to be written in English, but this is not always the case. The Twitter API provides information regarding the language of a given tweet, but many non-English tweets are labelled as English. We therefore use a state-of-the-art automatic language identification tool (Lui and Baldwin, 2011) to identify and remove non-English tweets.

In the following section we discuss the results of applying this method for identifying lexical blends from Twitter.

5. Results

We run a program implementing the method described in Section 4 for five weeks from late September to early November 2011. Some statistics about the number of tweets processed are presented in Table 2. (The total number of tweets is estimated based on Twitter's (2011) claims about the rate of tweets.) English tweets containing both a regular expression indicating that a word is new and a candidate blend are clearly rather rare; the frequency of such tweets is roughly 0.25 per million. Nevertheless, given the massive volume of data available through Twitter, we find roughly 49 such tweets per day. At this rate we are not overwhelmed by the number of hits, but still have enough data to work with that the findings are interesting. Among the 1717 tweets containing a candidate blend, there are 976 unique candidate blends.

Table 2. The number of tweets at various stages of processing.

Number of tweets	
Total (estimated)	7 000 000 000
Matching a regular expression	346 140
English and matching a regular expression	336 399
Containing a candidate blend	1717

To determine whether the candidate blends identified by our system are indeed blends, we annotate a random sample of 100 candidates. Based only on their usage in the tweets in which they were identified as candidates, 57 of the candidate blends are judged to be blends and have an interpretation corresponding to their respective candidate source words. It is unclear what level of precision is required for this method to be a practical lexicographical tool; however, Church (2010) notes that in a project on updating a thesaurus, an automatically-generated list of candidates having a precision of 10% was viewed as successful. Although requirements certainly vary from project to project, 57% precision is likely high enough to be useful in at least some cases, and in particular, our application of building a dataset of lexical blends.

We now consider how widely-used the candidates identified by our proposed method are. Although nonce formations and blends with very limited usage might be important to include in a dataset of blends intended for some purposes — such as documenting the full range of creativity observed in lexical blending — we are also interested in identifying which blends are in regular usage. The frequency of a term is an important indication of its salience, but the diversity of its usages and time span over which it has been observed are also important factors (e.g., Sheidlower, 1995; Metcalf, 2002). In this preliminary analysis we identify candidate blends that — in a sample of approximately 750 million tweets collected from September to November 2011 — are used in English tweets by at least 5 different users over a time span of at least 2 weeks. (The criteria of 5 different users and a 2 week time

period were chosen for this preliminary analysis, but could easily be set as appropriate for future analyses.) 453 of the 976 candidate blends meet these criteria. (Of the 57 items judged to be blends in the analysis in the previous paragraph, 20 meet these usage criteria.) This encouraging finding indicates that many of the candidates identified by our proposed method are not nonce formations, and might be (or might have been) in somewhat regular usage. Examples of well known blends identified by our method and meeting these usage criteria include *glocal* (*global* + *local*), *jeggings* (*jeans* + *leggings*), and *staycation* (*stay-at-home* + *vacation*).

One limitation of these usage heuristics is that a wordform for a blend might also have non-blend usages. For example, *stilly* is identified as a blend of *stupid* and *silly*, but *stilly* is also used to refer to Stillwater, Oklahoma, and this sense of *stilly* appears to be much more frequent. Blends which are potentially misspellings of other words pose similar problems. For example, *clearned* is found to be a blend of *cleared* and *cleaned*, but it is difficult to judge whether instances of *clearned* correspond to this blend interpretation, or simply to *cleaned* or *cleared* spelt in a non-standard way. This type of problem is rather common given that Twitter contains many such non-standard forms.

Furthermore, in a small number of cases we identify two types corresponding to the same lemma; for example, we find both *twatch* (*Twitter* + *watch*) and *twatching* (*Twitter* + *watching*). Similar challenges have been previously noted (e.g., Barnhart, 1985), and although additional processing (e.g., applying a lemmatiser) could potentially address some of these issues, further lexicographical analysis is required to determine which of these items are regularly used as blends. Nevertheless, such usage heuristics are useful for selecting candidates for further analysis.

We now examine a small number of candidate blends in more detail. For this analysis we focus on blends that appear to be in usage on Twitter, but that are not recorded in dictionaries. 79 of the 976 candidate blends meet the above usage criteria, and are not found in any dictionary searched by OneLook (http://www.onelook.com/). OneLook provides information about which of several online dictionaries, including UrbanDictionary (http://www.urbandictionary.com/), have an entry for a given headword; nevertheless, the information provided by OneLook for UrbanDictionary appears to be incomplete, so we also independently search UrbanDictionary. Although the contributors to UrbanDictionary do not necessarily adhere to any lexicographical principles, we include it in our search only to focus on the least-documented candidate blends.

We annotate the 79 candidate blends as to whether they are blends in the same manner as before (i.e., based only on the tweets in which they are identified as candidates) and find 30 to be blends of their candidate source words. We examined concordance lines from Twitter for each of these 30 blends and chose 10 to discuss further below.

- *Crasins (cranberries + raisins):* This blend refers to dried cranberries, which are made through a process similar to that which is used to make raisins from grapes. The term *Craisins* is a registered trademark, and appears to be more frequent on Twitter, but *crasins* is also in regular usage.

- *Dalaric (Damon + Alaric):* Damon Salvatore and Alaric Saltzman are fictional characters on the television show *The Vampire Diaries*, with *Dalaric* used to refer to the pair of characters or their relationship. Blends are often formed from the names of fictional characters, particularly in fan fiction, with examples related to *Harry Potter* including *Dramione* (*Draco Malfoy* + *Hermione Granger*) and *Snarry* (*Severus Snape* + *Harry Potter*). Blends of names are also common outside of fiction, with celebrity examples including *Brangelina* (Brad Pitt and Angelina Jolie) and *Bennifer* (Ben Affleck and Jennifer Lopez).

- *Julielmo* (*Julie* + *Elmo*): This is a blend of the names of two musicians, Julie Anne San Jose and Elmo Magalona.

- *Mocial* (*mobile* + *social*): This blend refers to the combination of mobile devices and social media, for example, 'We're gearing up for Internet Retailing next week! Read @<USER> thoughts on mocial marketing before the show: <URL> '

- *Neature* (*neat* + *nature*): This blend is often used in the expression *neature walk*, for example, 'Out on a neature walk through the Arboretum. What a beaut of a day! @<USER>' . *Neature Walk* is the title of a series of popular online videos.

- *Piloga* (*pilates* + *yoga*): This blend refers to a type of exercise that combines pilates and yoga. Other blends are also used to refer to combinations of these exercises, for example, *yogalates*.

- *Raincouver* (*rain* + *Vancouver*): This blend is often used as a hashtag (e.g., *#raincouver*), but is also common in unmarked contexts, for example, 'Oh no, Raincouver has reared its ugly head. It is November after all. Don't forget your umbrella today!' Place names are often used as source words in blends, with other examples including *Winterpeg* (*winter* + *Winnipeg*) and *perthonality* (*personality* + *Perth*).

- *Sockos* (*socks* + *Chacos*): This blend refers to wearing socks with Chacos, a brand of sandal. This blend appears to be the least frequent of those presented in this section, but we include it here because we are still able to find numerous usages that seem to correspond to it, for example, 'Oh sockos... so comfy... idk y its such a fashion fo-pah!!!' and 'RT @<USER>: Wore Sockos for the first time tonight.. Guess I can't make fun of @<USER> and @<USER> anymore cause they fe ...'

- *Twiple* (*Twitter* + *people*): This blend is often used in the context of greeting or parting, for example, 'Good afternoon twiple.....very very late start today but a good one' and 'Ok, time to sleep... Nite twiple'.

- *Twittership* (*Twitter* + *friendship*/*relationship*): This blend is commonly used on Twitter, for example, '@<USER> My hockey team? That's too personal. At this stage. Of our twittership.' An alternative, non-blend analysis of *Twittership* would be *Twitter* combined with the suffix *-ship*, influenced by words such as *sistership*, and *broship*. Furthermore, this lemma also appears to have a second but less frequent sense, not corresponding to a lexical blend, patterned on words such as *readership* and *viewership*, for example, 'Morning twittership. Another day to make a difference, how will you spend it?'

6. Discussion

The method we propose for identifying lexical blends involves two key steps: 1.) finding tweets matching regular expressions that indicate that a word is new, and 2.) finding candidate lexical blends amongst those tweets. In this section we consider the contribution of these individual steps. We begin by annotating a random sample of 100 of the 336 399 English tweets matching a regular expression (see Table 2). Amongst these tweets we find six usages

of lexical blends. These figures are not directly comparable to those of the previous section (in particular because the judgements in the present section are for tweets, not candidate blends, and do not consider the blends' source words) but clearly the rate of tweets containing a blend amongst tweets just matching a regular expression is relatively low.

We now consider the number of blends found amongst those tweets which include a candidate blend, but which do not necessarily match one of the 29 regular expressions. In this case we are no longer interested in specific keywords and so cannot use the Twitter API as we have until this point, that is, to only obtain those tweets matching particular keywords. Instead, we must retrieve and process a sample of all tweets from the Twitter API. By default Twitter provides access to a random sample of approximately 1% of all tweets (known as the "Spritzer"), but grants access to an approximately 10% sample, referred to as the "Gardenhose", upon request in some cases. For this analysis we obtained access to the Gardenhose, and retrieved all tweets from this source for 1 day during the period over which we collected the data for this study. This amounts to roughly 19 million tweets of which approximately 12 million are classified as English according to our statistical language identification tool. We find 2291 unique candidate blends amongst these English tweets. We randomly select 100 of these candidates to analyze. Based only on their usage in the tweets in which they were identified as candidates, we find 14 to be blends of their respective candidate source words. This indicates that the rate of lexical blends is higher amongst candidates obtained from tweets containing a candidate blend and a regular expression indicating that a word is new, than amongst candidates from tweets just containing a candidate blend; however, it does appear that a greater number of blends can be found by processing the Gardenhose than by our proposed 2-step approach, but this comes at the cost of manually examining a greater number of noisy candidates. Furthermore, our proposed approach requires only the default access level for the Twitter API, and has substantially lower computational cost than an approach which processes the Gardenhose.

7. Conclusions

This paper presented a simple computational method for identifying English lexical blends by exploiting the massive amount of text available on Twitter. When applied to five weeks of data, the proposed method identified 976 candidate blends; analysis of a sample of these candidates indicated that 57% are blends. We further discussed a small number of blends identified by our method that are not recorded in dictionaries, but that were found to be in regular usage on Twitter.

To date we have run our proposed method for identifying English lexical blends for approximately 5 months and have identified over 7000 unique candidate blends. In the future we plan to analyze many more of the candidates identified by our system to produce a dataset of blends, which we intend to make publicly available. Given the large (and growing) number of candidate blends to analyze, we intend to make use of statistical computational methods, similar to those of Cook and Stevenson (2010), to rank the candidates according to their plausibility as blends.

To encourage further research on lexical blends and the use of Twitter for lexicography, and to make it easier to verify the findings of this paper, computer programs implementing the presented approach to identifying lexical blends (the programs discussed in Section 4) have been made available on the author's website (www.cs.toronto.edu/~pcook).

Note

[1] Thanks to Jane Solomon, the members of the University of Melbourne Language Technology Group, and the members of the NICTA BioTALA project for their feedback on this work. Additional thanks to Aaron Harwood, Shanika Karunasekera, and Masud Moshtaghi of the University of Melbourne Department of Computing and Information Systems for supporting this research by providing access to, and help in processing, data collected from Twitter.

References

A. Dictionaries

Barrett, G. 2006. *The Official Dictionary of Unofficial English*. New York: McGraw-Hill.

B. Other literature

Algeo, J. (ed.) 1991. *Fifty Years Among the New Words*. Cambridge: Cambridge University Press.

Barnhart, D. K. 1985. 'Prizes and Pitfalls of Computerized Searching for New Words for Dictionaries'. *Dictionaries*, 7: 253–260.

Bird, S., E. Loper and E. Klein 2009. *Natural Language Processing with Python.* Sebastopol: O'Reilly Media Inc.

Church, K.W. 2010. 'More is More.' In G-M. de Schryver, (ed.), *A Way with Words: Recent Advances in Lexical Theory and Analysis: A Festschrift for Patrick Hanks*. Kampala: Menha Publishers,135–141.

Cook, P. 2010. *Exploiting Linguistic Knowledge to Infer Properties of Neologisms*. PhD Thesis, University of Toronto.

Cook, P. and S. Stevenson 2010. 'Automatically Identifying the Source Words of Lexical Blends in English.' *Computational Linguistics*, 36(1): 129–149.

Kaufmann, J. and J. Kalita 2010. 'Syntactic Normalization of Twitter Messages.' In *International Conference on Natural Language Processing*, Kharagpur: ICON 2010.

Lui, M. and T. Baldwin 2011. 'Cross-domain Feature Selection for Language Identification.' In H. Wang and D. Yarowsky (eds.), *Proceedings of the Fifth International Joint Conference on Natural Language Processing (IJCNLP 2011)*. Chiang Mai: Asian Federation of Natural Language Processing, 553–561.

Metcalf, A. 2002. *Predicting New Words*. Boston: Houghton Mifflin Company.

Porter, M. F. 1980. 'An Algorithm for Suffix Stripping.' *Program*, 14(3): 130–137.

Renner, V., F. Maniez, P. J. L. Arnaud (eds.) (forthcoming). *Cross-disciplinary Perspectives on Lexical Blending*. Berlin / New York: Mounton De Gruyter.

Sheidlower, J. T. 1995. 'Principles for the Inclusion of New Words in College Dictionaries.' *Dictionaries* 16: 33–44.

Twitter. 13 October 2011. *200 Million Tweets per Day*. http://blog.twitter.com/2011/06/200-million-tweets-per-day.html.