Twitter and the Geo-linguistic fingerprint

Eitan Altman

Twitter and the Geo-linguistic fingerprint

Eitan Altman
INRIA Sophia-Antipolis, 2004 Route des Lucioles, 06902 Sophia-Antipolis Cedex, France
email: Eitan.Altmaninria.fr

February 28, 2012

Abstract

Having access to content of messages sent by some given group of subscribers of a social network may be used to identify (and quantify) some features of that group. The feature can stand for the level of interest in some event or product, or for the popularity of some idea, or a musical hit or of a political figure. The feature can also stand for the way the written language is used and transformed, the way words are spelled and grammar is used. In this paper we shall be interested in identifying features of groups of subscribers that have their geographic location and their language in common. We develop a methodology that allows one to perform such a study using a statistical tool which is freely available, and which makes use of a part of all tweets which twitter makes available for free over the Internet. The methodology is based on the fact that one can differentiate among some geographic areas according to the activity pattern of tweets during the time of the day. We present an application of this methodology to the study of new spellings or of new words created in twitter messages.

1 Introduction

Unlike many other social networks whose business model is mainly based on offering advertisements, twitter makes money by selling content: the content of a large portion of transmitted messages is sold to interested companies. One can buy almost all the content for around thirty thousand dollars a month. One can receive smaller portions for lower prices. A small portion of around 1% is made available for free. The fact that such a huge amount of messages is made available makes twitter attractive as a tool for learning about opinions in a large population. Twitter can serve as an alternative to opinion polls for market analysis not only in the context of selling goods but also for opinion trends analysis such as election campaigns [1, 2]. Twitter allows to access some information for free through different APIs (Application Program Interface).

The methodology is based on the fact that one can differentiate among some geographic areas according to the activity pattern of tweets during the time of the day. More precisely, we make use of the fact that the amount of messages generated by subscribers at a given location changes during the time of the day in a periodic way which may differ from one region to another. For example, this activity is much lower when most people in that region are asleep late at night.

We apply this methodology to the study of new spellings or of new words created in twitter messages.

We note that there are other ways of obtaining geo-localisation of messages as well as the identification of language in which they are written, based on information that are available in some of tweets. The use of such information would require the user to have software tools that are not available on the Internet for free public use. We thus decided to focus in this paper on a methodology that can be widely used relying on the "trendistic" API (availble for free use on the Internet, see http://trendistic.indextank.com/).

2 Periodograms of daily activity: a geo-linguistic fingerprint

Figure 1 displays the frequency of appearance of the words "to, the, el, y, a, i" over a period of a month. The frequency of each of these words has a periodic behavior where the period corresponds to one day. We also observe that the words "To", "I" and "The" have a very similar wave form, and so do the words "El" and "Y". The word "A" has a distinct wave form differenet from the other two. The first group contains words that are frequently used in English, where as the second group corresponds to words that appear frequently in Spanish. The word "A" appears frequently in many languages (e.g. English, Spanish, French). The word Y appears also in French but its frequency there is much smaller. We conclude that words that are typical to one specific language have a common pattern, which we call a "fingerprint" of the language.

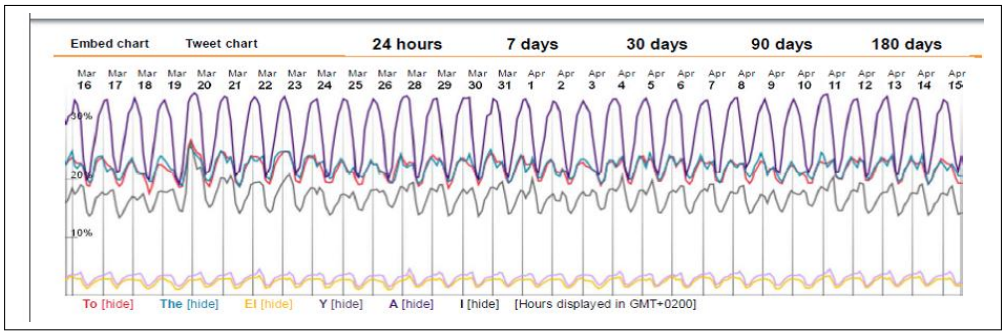

Figure 1: The frequency of appearance of the words "to, the, el, y, a, i"

Figure 2 presents a typical german finger print. The same daily period is seen to be common to three different words that are very common in German.

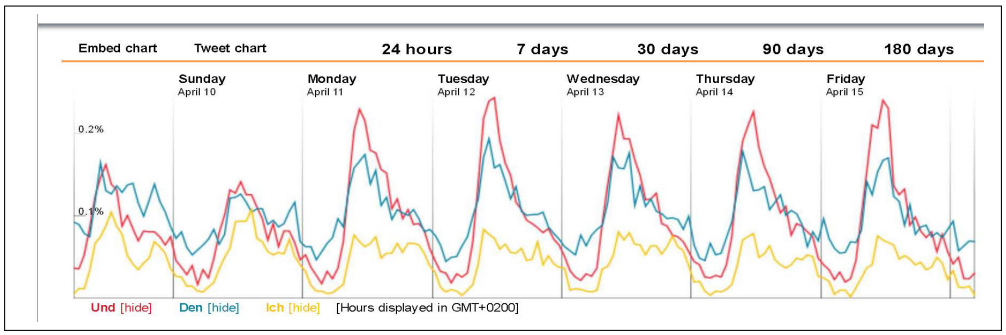

Figure 2: The frequency of appearance of several german words during 7 days. A very clear common periodic daily pattern appears.

Next we give an example of several spanish words, see Figure 3. There are two exceptions. We included an English word, "the", which is the most popular word in the figure. It appeared in around 9% of all tweets. It indeed has a completely different periodic pattern. The secondn exception is the word "La" which frequently appears not only in Spanish but also in French and Italian. Nevnerthless, unlike "and", we observe much reseemblence to the pattern of Spanish words. A possible explanation could be that there are significantly more tweets in Spanish than in French and Italian. Therefore the periodogram of "La" is closer to the spanish even if spanish and french words had quite different periods.

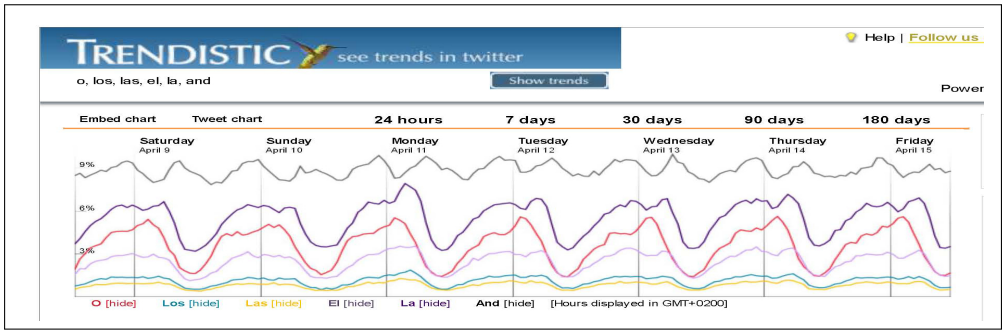

Figure 3: The frequency of appearance of several spanish words during 7 days. A very clear common periodic daily pattern appears and is compared to non-spanish words

The reason that each language has its own fingerprint could be

- The fact that each language has its own geographic distribution, and thus a different time-zone distribution.
- The habits related to working hours, eating hours etc may differ from one community to another, and these habits may imply different distribution of tweeting times.

Can we check which of the above is more pertinent? Observe in Figure 4 the frequency of appearance of the words "une", "della", "der". These three words correspond to articles in French, Italian and German. We see that the periodic frequency pattern of the three words is very similar. These three languages correspond are mainly spoken mainly in Europ, and the time zone in which they are spoken is the same. It thus seems that the geographic location plays a major role so that similar geographic location indeed gives similar fingerprints.

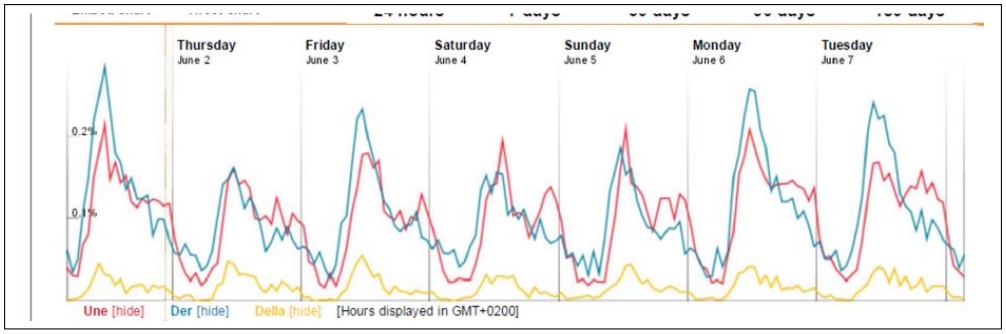

Figure 4: The frequency of appearance of the words "une", "della", "der"

We next compare the Spanish word "Todas" with the French word "et". The Spanish one is seen to be shifted with respect to the French word by around 6 hours. For example its lowest activity during the day appears around 6 hours later than that of the French word. This suggests that most tweets in spanish originate in Latin America which has a time difference of 6 hours or more with respect to France.

3 More detailed geo-linguistic fingerprints

Daily periodograms can be made more selective so as to restrict to a subregion in which a language is spoken. As an example, we compare tweets with the Spanish words "computadora" and "ordenador". Both mean "computer", but the first is used in Spain and the secon in Latin America. The corresponding peroiodograms appear in Figure 5. We see that the term used in Spain has its minimal appearence around 8 hours before the Latin American one.

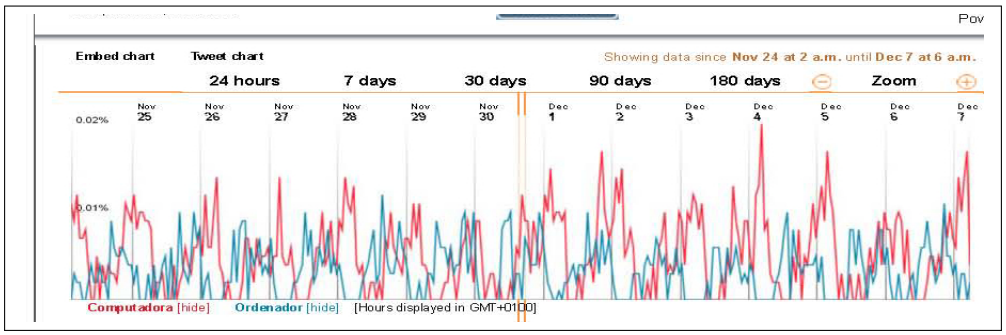

Figure 5: The frequency of appearance of a word in Spanish from Latin America and that from Spain

We note that the average daily number of tweets in which "ordenador" appears is around 2/3 the one corresponding to "computadoras". Does this suggest that the fraction of spanish tweets originatinng from Spain is close to that originating from Latin America? To answer this question, we may wish to compare also other words, or in contrast, to see how the relative frequencies behave in other contexts. When comparing the number of appearence of these words over the whole Internet, by using fightgoogle, we obtained (on Dec. 7, 2011) the figures: 4,580,000 for "computadora", and 7,150,000 for "ordenador".

Next we shall differentiate between the periodograms of the American versus the British versions of English. We do so by comparing the fraction of tweets containing "realize" (American version) and "realise" (British version) as a function of time, as is seen in Figure 6.

We see clearly that the minimum daily activity of the American word occurs around 6 hours later than the British one.

4 Twinglish and other languages

"My son" in Spanish appears in Twitter often as "mijo" which is an abreviation of the two words "mi hijo". Figure 7 shows the daily pattern of the use of the word. All apearences of the word which we observed were indeed in Spanish. There is a clear inactivity period that corresponds to around 8am in French time. We conclde that the term "mijo" probably

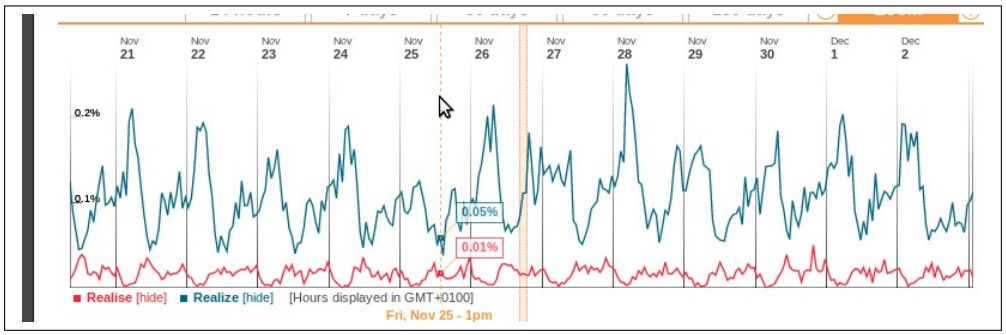

Figure 6: The frequency of appearance of the words "realise" and "realize"

originates from the west part of Latin America. Similar behavior characterizes the word "porfa" whose periodogram is given in Figure 8. This is a way of shortening the word "please" in Spanish, which is written as "por favor".

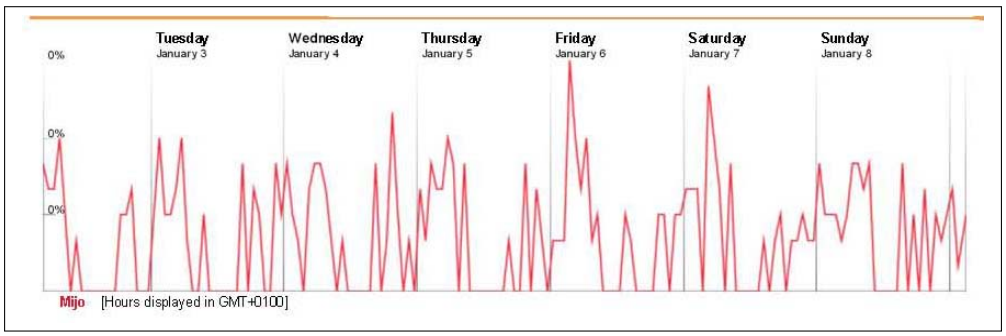

Figure 7: The frequency of appearance of "mijo"

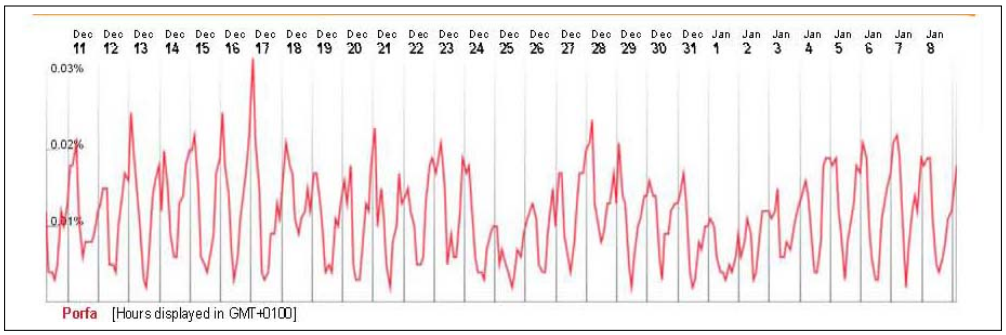

Figure 8: The frequency of appearance of "porfa"

In contrast, the word "xk" which means in tweepanish "because" or "why" and is pronounced "porque" has no inactivity periods, see Figure 9. "xk" is much less localized and is probably used both in latin America and in Spain. Note that the translation of "por" using "x" is due to the interpretation of x as multiplication, which is pronounced as "por".

We next observe the evolution of the word "xo".

The online urban dictionary http://www.urbandictionary.com/ says that x means kiss and o means hugs. xoxo then means "kisses and hugs". We found out that in Spanish "xo"

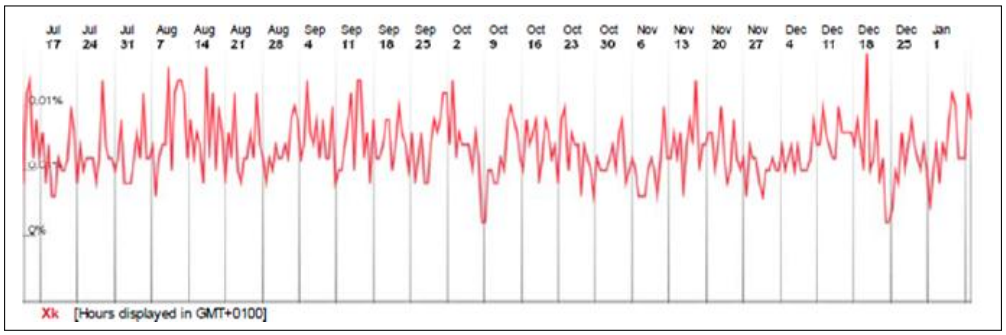

Figure 9: The frequency of appearance of "xk"

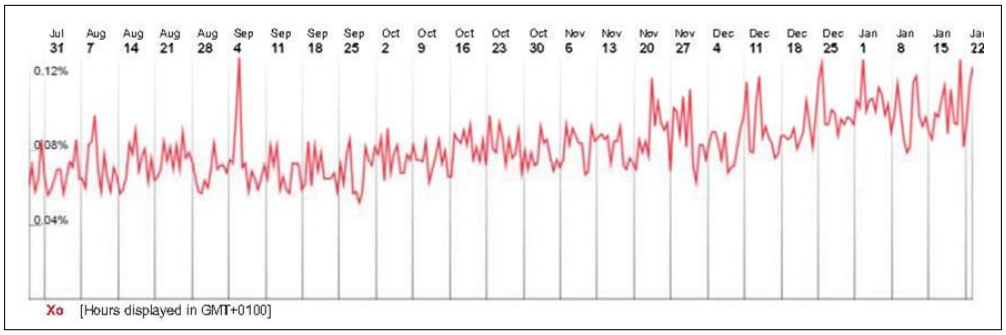

Figure 10: The frequency of appearance of "xo"

is also used to say "pero" ("but" in English), where the explanation for the use of x is as in xk.

When working with trendistic, we can use the ratio between the maximum and minimum of the activity level during a day as a measure of its locality. We shall say that a term is well localized if this ratio is larger than 2.

5 The Spanish word Porque

We discuss in some more details the spelling we find in twitter for the words "porque" and "because".

We already mentionned the spelling "xk" for "porque". We found many other spellings. We list them along with the number of tweets in which they appear averaged over the six months period of beginning of Aug 2011 - end January 2012.

We tried also the following spellings: "porque" (0.5%), "xq" (0.07%), "porq" (0.04%), "xk" (0.012%). The frequency of their appearence in twitter is depicted in Figures 11 and 12.

Other spelling had too few occurrences and trendistics gave the message "There is too little data for a full chart so we are showing only recent activity". These spellings are "podque", "podq", "podk". They are obtained by replacing the "r" by "d" in the word "porque" and then, for the two last spellings, "que" is abreviated. Such a replacement is a typical childish way of speaking spanish, as many children have difficulties to pronounce the r and replace it the by d.

We found no tweets with the spelling "xque". The spelling "pork" appears, but most tweets with this spelling correspond to the English word "pork".

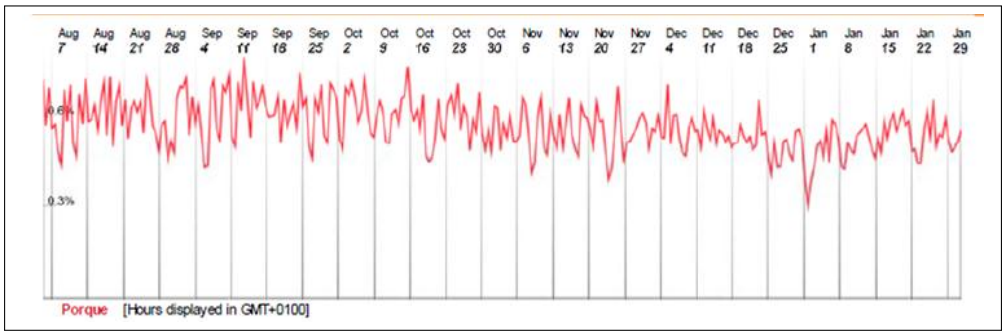

Figure 11: The frequency of appearance of "porque"

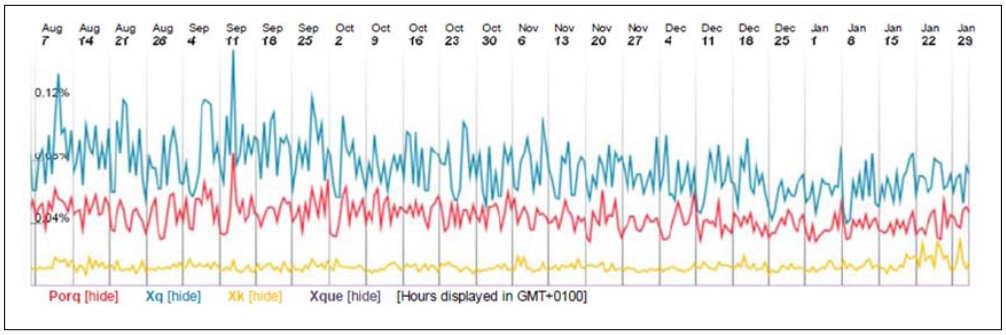

Figure 12: The frequency of appearance of other spellings of "porque"

6 The English word "Because"

The word "because" appears in twitter with a large number of possible spellings.

Two short forms appear frequently in twitter: "cuz" (around 0.2% of tweets) and "coz" (around 0.02% of tweets). Twinglish thus allows us not only to recognize the word but also to hear it, and hence distinguish between the American and British accents. In Figure 13 we observe the periodograms of both. We see that "cuz" and "coz" have exactly the opposit activity profile: the minimum activity of "coz" are during night hours in Europe where as those of "cuz" are in night time in USA and Canada. The maximum daily activity of "coz" is during day time in Europe where as "cuz" has its maximum activity at day time in America. The periodogram of both words show very well localization: the ratio between the peak and the minimum activity is around 5 for both "cuz" and "coz".

We also find the spelling "cus" and "cos" as seen in Figure 14-15. Again, the spelling is seen to correspond to the accent. The geo-linguistical finger print of "coz" is seen to be the same as "cos" (Fig 14). They both correspond to the UK where the second vower of "because" sounds like "o", as opposed to the American pronounciation that sounds like "u" which we find in "cuz" and "cus".

When comparing the two "American" spellings "cuz" and "cus", we see that there is a

Spelling:	because	cuz	cos	coz	cus	bcuz	cz	becuz
% of tweets in which it appears:	0.7	0.2	0.03	0.02	0.02	0.015	0.012	0.011

Table 1: The most popullar spellings of "because" in Twitter

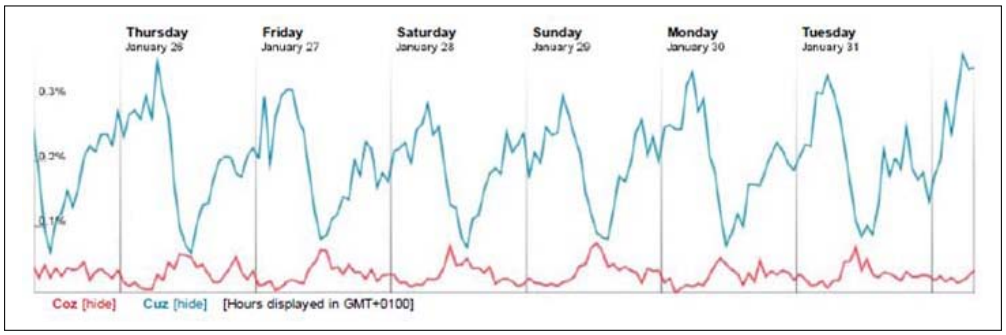

Figure 13: The frequency of appearance of popular spellings of "because"

very clear preference to "cuz" where as the British seem quite indifferent between the two British spellings "cos" and "coz". The preference of the version with "z" in the America is in line with the fact that there have been already much before twitter differences between UK and USA with respect to the use of s versus z.

Further shortning of "coz" and "cuz" by elliminating the vowel is possible but it did not seem appealling to Twitternauts. We have not found "because" written as "cs". It appearead however as "cz", four times less frequently than "coz". From its periodogram in Figure 16, "cz" is seen to be very localized and it corresponds to the same activity period as that of "coz". We conclude that the use of "cz" is restricted to Twitternauts from UK.

Two other spelling, "bcuz" and "becuz", appear with lower activity Their periodogram in Figure 17 shows activity periods that correspond to in America. We again have high degree of localisation. We did not find tweets with the spelling "bcz" or "bcos".

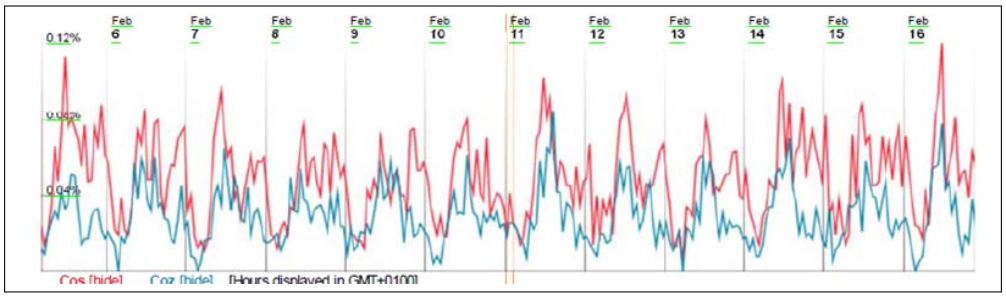

Figure 14: The frequency of appearance of the spellings of "cos" and "coz" of "because"

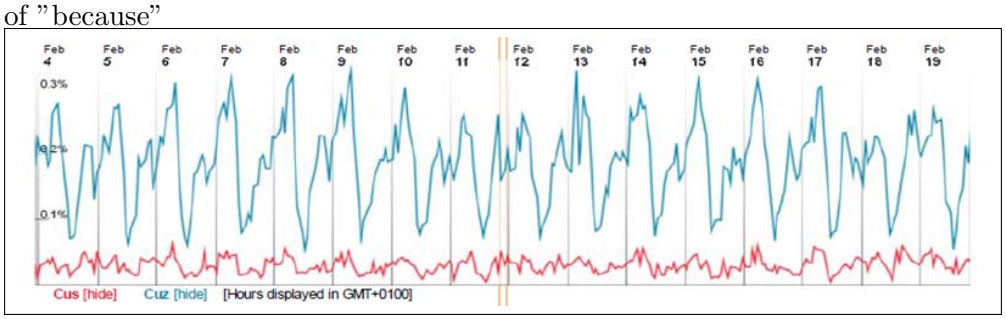

Figure 15: The frequency of appearance of the spellings "cus" and "cuz" of "because"

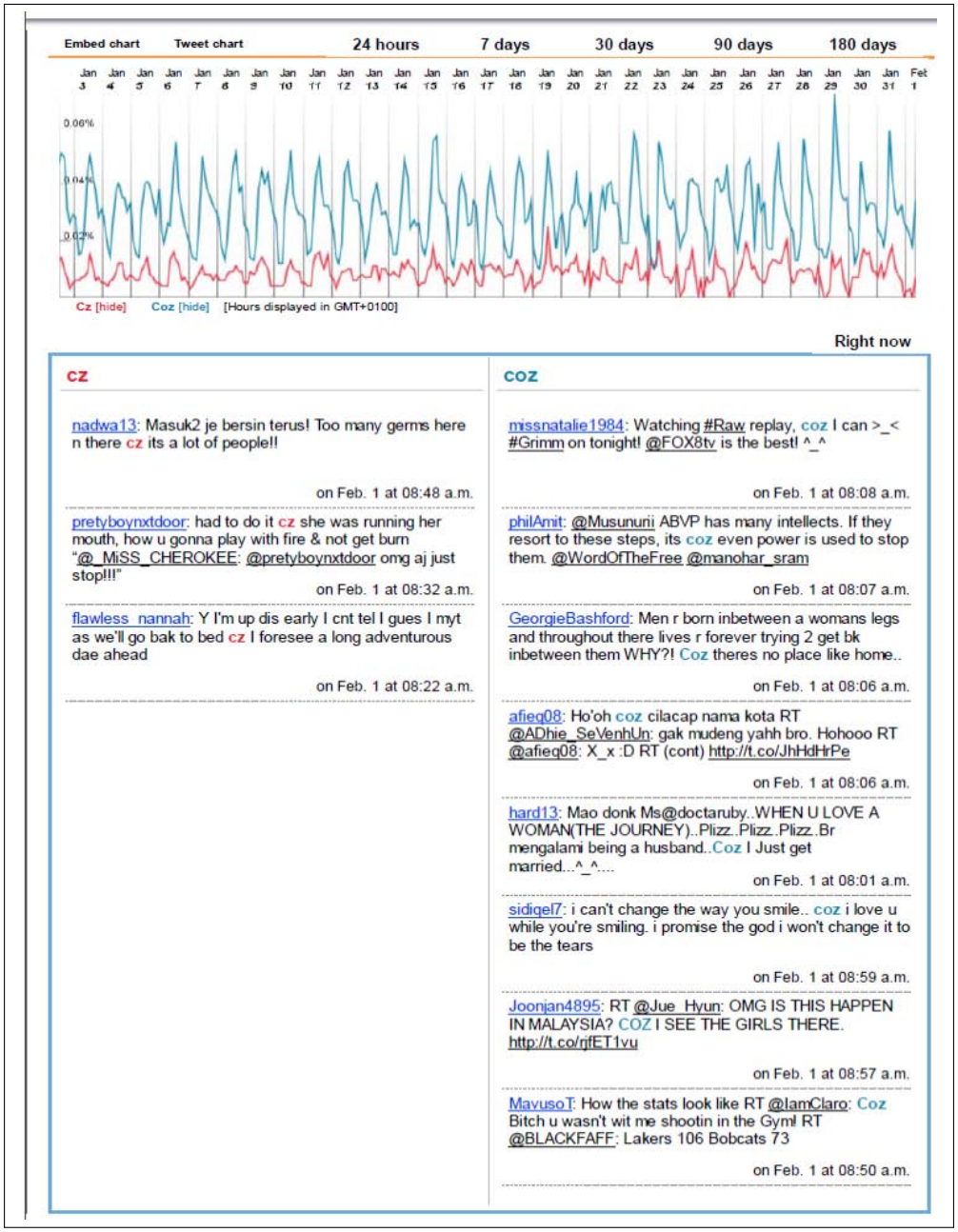

Figure 16: The frequency of appearance of other spellings of "cz"

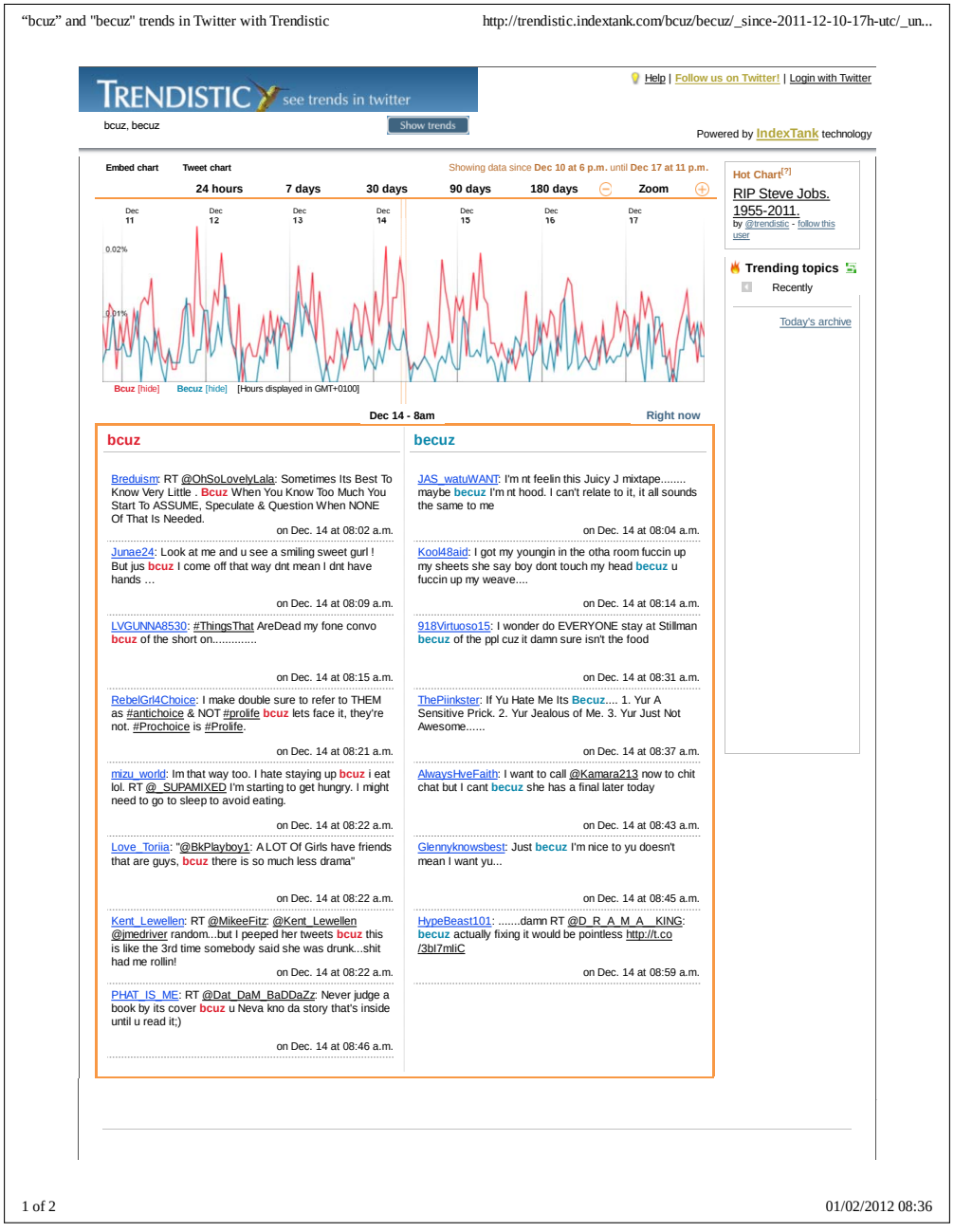

Figure 17: The frequency of appearance of the spelling "bcuz" and "becuz"

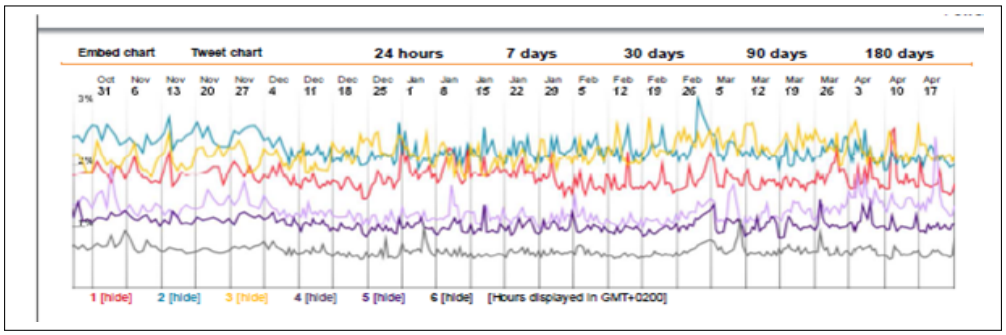

Figure 18: The relative frequencies of the integers 1, 2, 3, 4, 5, 6

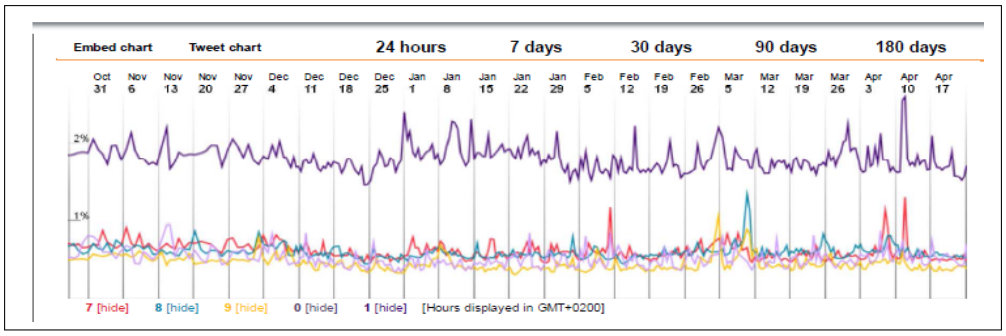

Figure 19: The relative frequencies of the integers 7, 8, 9, 0, 1

7 Numbers

We tried the following experiment. We compared the daily frequency of the appearance of the integers 1, 2, 3, ... in Twitter messages. We expected to obtain variations around some common average: we did not expect one number to appear more frequently than another in the long run. Figure 18 shows the measured frequencies.

We observe significant differences between the frequencies of appearance of various integers. What is the reason for that? Looking into the messages themselves, one immediately observes that through the way of pronouncing them, integer numbers have other meanings as well. Integer numbers are then used so as to shorten the number of characters needed for transitting messages. This is extremely useful in Twitter since it

In particular

- 4 is pronounced the same as **for**. Therefore, 4u can be used to write the two words "for you" (which contain seven characters) in two characters only.
- 2 is pronounced like "to". It thus allows again to reduce the number of characters. E.g., writing 2u allows us to express "to you" in 2 characters.
- 3 is used frequently combined with < which gives <3 This can be interpreted as a kiss or a heart, see Fig 20.

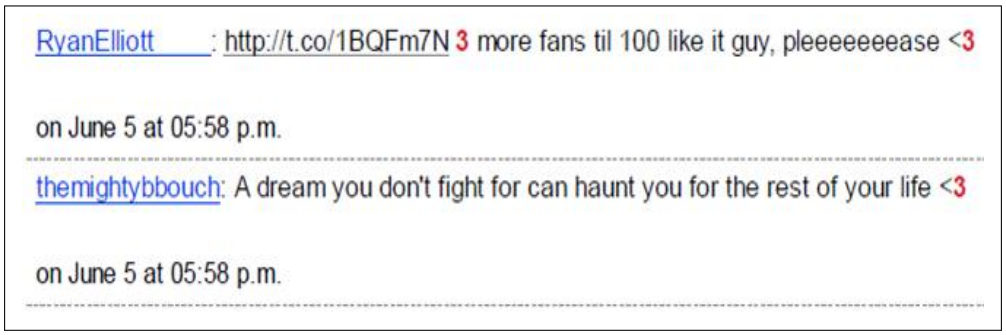

Figure 20: Examples of tweets using "3"

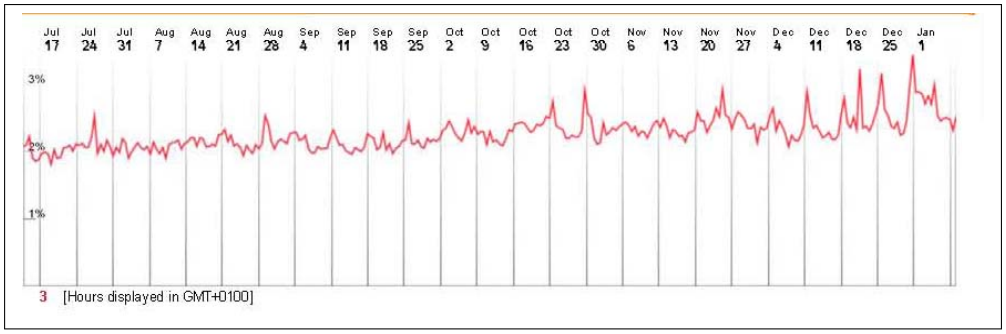

Figure 21: The time evolution of "3"

7.1 On 3 and its Evolution

The frequency of the use of 3 increases rapidly as is seen in Fig. 21. Since mid july 2011 till mid January 2012 its popularity increased from arounud 2% of the tweets to 2.7%. Its peak is achieved on Dec. 31 with a popularity of 3.42%. During this period, the other numbers showed negligible variations in popularity.

The number 3 has several other meanings: 3Q means "thank you" in chinese: 3 is pronounced as "SUN" and "Q" is pronounced as "Queue" so that 3Q together immitates the sound of the English word "Thank you".

The digits 3, 5, 7, and 9 are used also as letters in Arabic when using a latin keyboard to write arabic. They correspond to the arabic letters "Ain", "Kha", "Cha", "Ka", respectively.

Some use digits also in coding latin letters: 0 instead of "o", 4 i.o. A, 3 i.o. e, 1 i.e. i, 5 i.o. s.

8 Conclusions

We have presented some examples that showed an amazingly fast evolution of the vocabilary over Twitter. The creation of new words is often explained by the advantages in writing shorter words: both the character limitation in twitter as well as the fact that many tweets are sent from cellular phones whose small keyboard is not as confortable as that of a laptop.

In the creation process of new spellings, alpha-numerical symbols often replace cylables according to

- (i) the phonetic sound that they are associated with. Examples are 3Q which is used in

Chinese as for "thank you". We call this an "audio association".

- (ii) the graphic form that they have. That symbol "¡3" is a "graphic association" of a heart or lips and is used for expressing affection. The number 7 has a form similar to that of the letter "cha" in Arabic and is thus used as such when an arabic keyboard is not available.

- (iii) Composition of associations: we saw that "xk" means "because" The "x" is pronounced "por" through a two step association: first a graphical association is used to transform "x" to "multiply", and then the audio association of "multiply", which is "por" in Spanish, is used.

The audio associations are often innexact. Here are some examples.

The letter "k" is pronounced as "ka" in Spanish so that "xk" sounds as "porqua" where as it is used in the meaning of "because" in Spanish, which sounds like "porque".

"k2" sounds as "KaDeu" in French and means a "present"; the pronounciation of "present" in French is, however, "KaDo".

The word "your" is often shortened to "yo".

It is not a surprise that there is a big tolerancne to such imprecisions, as we know of natural languages in which the vowels, altogether, do not appear in the written version (e.g. Hebrew or Arabic). Yet, although we see vowels appear often in an imprecise way in Spanish, English and French, the Twitterenauts do not seem eager to drop them completely (as we saw in the shortning of "because").

We saw that Twitternauts often convay the accents they use. This was the case of the word "because" whose twitter spelling spelling "cuz", "becuz" or "bcuz" suggest the USA accent whereas its spelling "coz" suggests the British one. We showed that this classification is confirmed with a high degree of localization obtained using the periodograms.

Further audible features of words appeared, e.g. in replacing the "r"s by "d"s in Spanish, as we saw in the word "porque". This feature also occurs in English, where the sound "th" in words such as "the", "this" and "that" is sometimes pronounceed as a "d". We illustrate this in Figure 22.

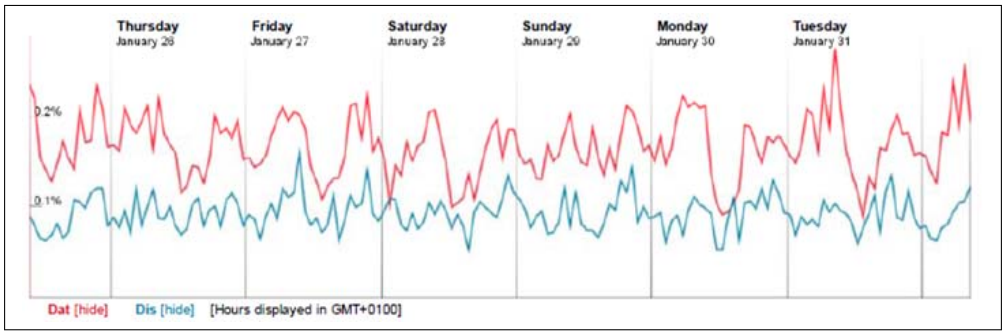

Figure 22: The words "this" and "that" spelled as "dis" and "dat"

References

[1] O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In Proc. 4th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM).

[2] Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Proc. 4th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM).