



**HAL**  
open science

# Geo-linguistic fingerprint and the evolution of languages in Twitter

Eitan Altman, Yonathan Portilla

► **To cite this version:**

Eitan Altman, Yonathan Portilla. Geo-linguistic fingerprint and the evolution of languages in Twitter. [Research Report] 2012, pp.14. hal-00674853v2

**HAL Id: hal-00674853**

**<https://inria.hal.science/hal-00674853v2>**

Submitted on 24 Apr 2012

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



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.

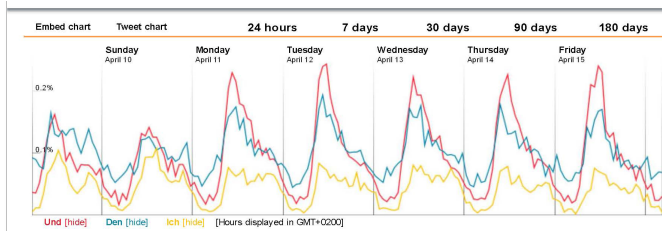


Fig. 2. The frequency of appearance of several german words during 7 days. A very clear common periodic daily pattern appears.

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 second exception is the word "La" which frequently appears not only in Spanish but also in French and Italian. Nevnertheless, unlike "and", we observe much resemblance 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.

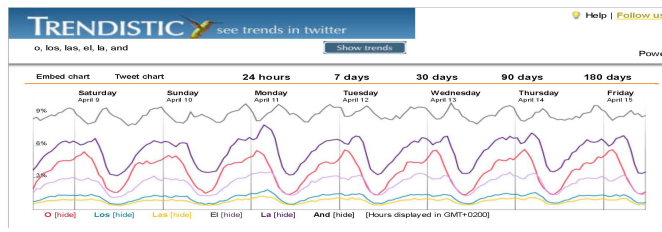


Fig. 3. The frequency of appearance of several spanish words during 7 days. A very clear common periodic daily pattern appears and is compared to non-spanish words

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.

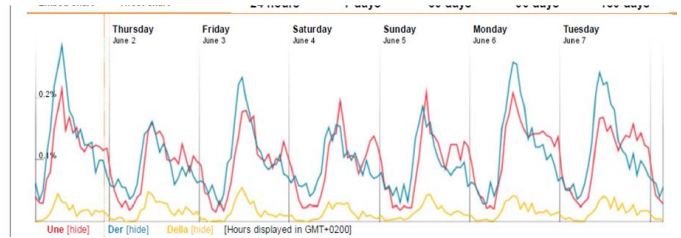


Fig. 4. The frequency of appearance of the words "une", "della", "der"

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.

### III. 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 Latin America the second in Spain. The corresponding peroidograms appear in Figure 5. We see that the term used in Spain has its minimal appearance around 8 hours before the Latin American one.

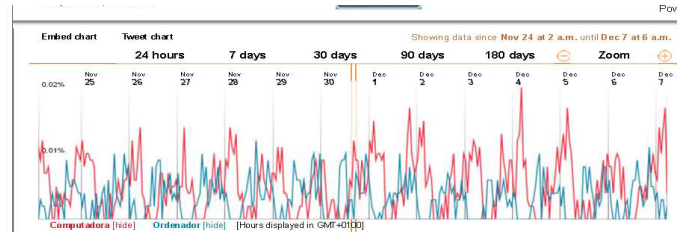


Fig. 5. The frequency of appearance of a word in Spanish from Latin America and that from Spain

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 originating 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 appearance 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

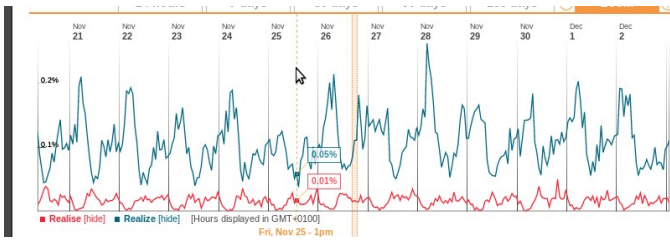


Fig. 6. The frequency of appearance of the words "realise" and "realize"

one.

#### IV. TWINGLISH AND OTHER LANGUAGES

We focus in this section on some spellings or forms of writing words that are typical for social media [3], [4], [5]. We highlight some geographic aspects related to these.

"My son" in Spanish appears in Twitter often as "mijo" which is an abbreviation of the two words "mi hijo". Figure 7 shows the daily pattern of the use of the word. All appearances 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 conclude that the term "mijo" probably 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".

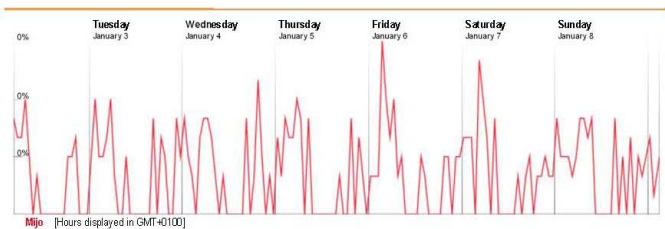


Fig. 7. The frequency of appearance of "mijo"

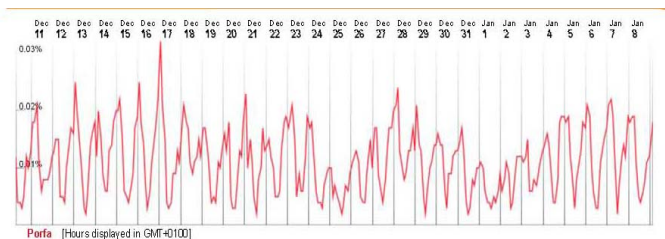


Fig. 8. The frequency of appearance of "porfa"

In contrast, the word "xk" which means in twitter-Spanish "because" or "why" [5] 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

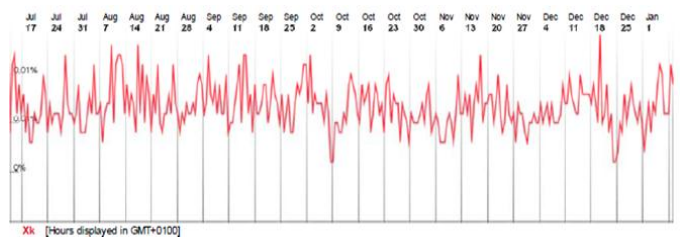


Fig. 9. The frequency of appearance of "xk"

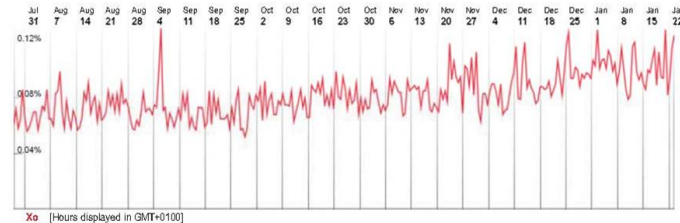


Fig. 10. The frequency of appearance of "xo"

and o means hugs. xoxo then means "kisses and hugs". We found out that in Spanish "xo" is also used to say "pero" ("but" in English), where the explanation for the use of *x* is as in *xk*.

In addition to spanish, *xk* is also used in the same meaning in Italian (for the Italian word "perche" meaning "why"), see [6]. Note that the first vowel is pronounced different than in Spanish.

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.

#### V. THE SPANISH WORD PORQUE

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

We already mentioned 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 appearance 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 abbreviated. 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".

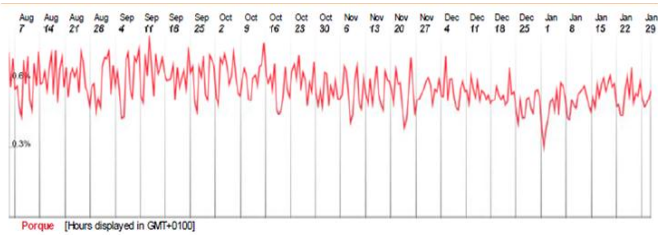


Fig. 11. The frequency of appearance of "porque"

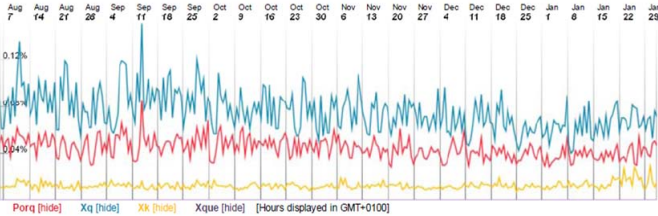


Fig. 12. The frequency of appearance of other spellings of "porque"

## VI. THE ENGLISH WORD "BECAUSE"

The word "because" appears in twitter with a large number of variations. In fact, the total fraction of tweets in which this word appears in a new form and/or spelling is larger than that corresponding to the original word. This is illustrated in Table I which provides the most frequent variations of "because" along with the number of tweets in which they appear (averaged over 300 samples) along with the standard deviation. For reference, the average fraction of tweets in which the word "because" appears without changes is 0.824% of the tweets with a standard deviation of 0.107%.

Two short forms appear frequently in twitter: "cuz" (around 0.2% of tweets) and "coz" (around 0.02% of tweets). Such shortening of words are called "clipping" in linguistic research [4]. Twinglish thus allows us not only to recognize the word but also to hear it, and hence distinguish between the American and British accents. (Note that the opposite happened with the word "xk" which means "why" both in Italian and in Spanish, but is pronounced differently in the two languages.)

In Figure 13 we observe the periodograms of both. We see that "cuz" and "coz" have exactly the opposite 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 vowel of "because" sounds like "o", as opposed to the American pronunciation that sounds like "u" which we find in "cuz" and "cus".

When comparing the two "American" spellings "cuz" and

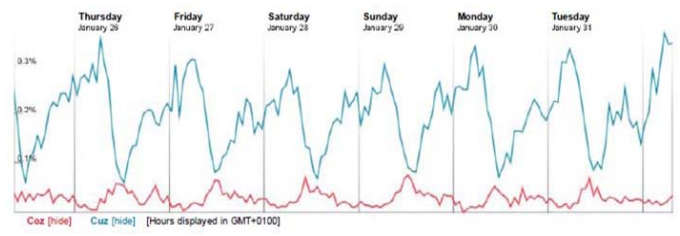


Fig. 13. The frequency of appearance of popular spellings of "because"

"cus", we see that there is a 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 eliminating the vowel is possible but it did not seem appealing to Twiternauts. We have not found "because" written as "cs". It appeared 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 Twiternauts 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".

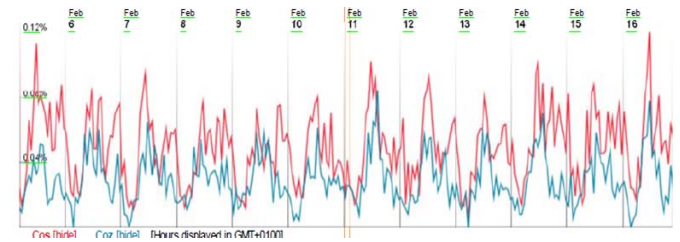


Fig. 14. The frequency of appearance of the spellings "cos" and "coz" of "because"

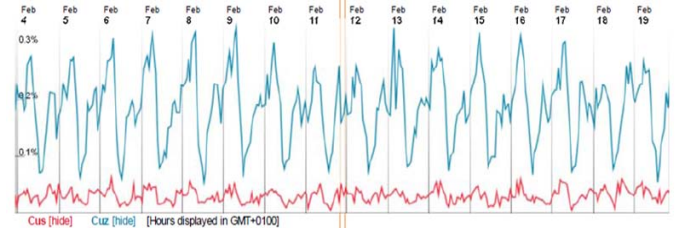


Fig. 15. The frequency of appearance of the spellings "cus" and "cuz" of "because"

**Remark 1.** Many of the spellings that we presented have other meanings than because. For example, *cos* is also the mathematical function "cosine", "cause" usually means "reason" and is also a verb, "cz" is used in other context for the Czech Republic (it has also other meanings). Thanks

Spelling:	because	bcuz	becuz	cause	cuz	cos	cus	coz	cz
% of tweets in which it appears:	0.824	0.0156	0.0116	0.505	0.187	0.0490	0.0304	0.0135	0.0128
Standard deviation	0.107	0.00350	0.00355	0.0678	0.011	0.0141	0.064	0.011	0.00355

TABLE I  
THE MOST POPULAR SPELLINGS OF "BECAUSE" IN TWITTER

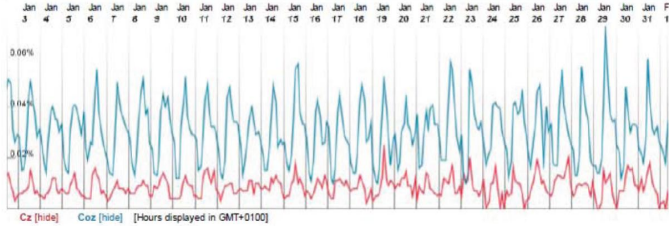


Fig. 16. The frequency of appearance of other spellings of "cz"

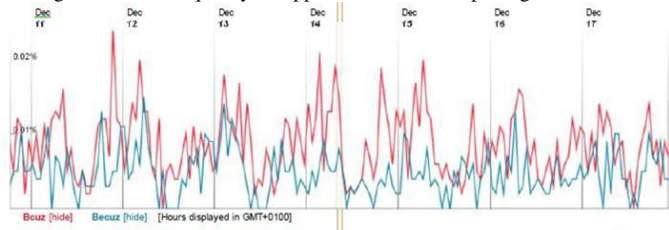


Fig. 17. The frequency of appearance of the spelling "bcuz" and "becuz"

to the snapshots of the contents of the tweets that trendistics provides, we were able to confirm that the above spellings are indeed used in twitter mainly in the sence of because. We shall later see that this is not the case in other electronic media.

## VII. COMBINATION OF DIFFERENT SPELLINGS

In view of the large number of different spelling of words, the natural question that arises is of what information is conveyed by a different spelling of a word. We already saw that the spelling can convey to which geo-linguistic group one belongs, e.g. whether the message is written by a British or an American English speaking person. The differentiation between these English versions that already existed since long ago has increased in tweeter, not only with respect to other media but also with respect to other social networks.

In this Section we examine other information and semantic content that spelling conveys. This has been motivated by a large number of messages in twitter in which we found various spellings of the same word in the same message. Is it intentional?

We focus on the word "because" in English and Spanish. Here are some examples of intentional use of different spellings or forms of the word.

We found different spellings appearing in the same sentence when comparing between two reasons for some action or some decision. Here are some examples that we found in tweets written in the end of March 2012 and the beginning of April.

- 1) 26 March, 2012: "laugh at u because ur different?? Laugh at them becuz they're the same!"

The example did not specify in what sense the one or the others are different or not. This is left to the imagination of the reader. My own preferred interpretation is that the difference is in the way of spelling "because", and hence the message advises one not to be too hurt if they laugh at you because you use a different spelling for writing "becuz", but rather to laugh at them becuz they all write becuz using the same spelling.

- 2) 26 March, 2012: if you haven't notice by now... i'm really bad at replying to ppl, not becuz im ignoring you, just because im a lazy person lol.
- 3) March 27th, 2012. BTW...that LOL was because of fun memories...not becuz of song that was tweeted by RT
- 4) 27 march, 2012: i cannot wait til spring break 2013, nt because of the trip but becuz on my way back ill be saying, this is my last time drivin to muncie :)
- 5) April 2nd, 2012: In class a professor called US a pure socialist country becuz it bails out all its companies, and Russia a capitalistic one because it wudnt
- 6) April 4th, 2012: U don't want to get to kno me becuz of something I did or somewhere I've been...u want to get to kno me just because of how I look..

The same structure appears also in Spanish, for example:

- 4th April, 2012: no llueve porque tu salgas del colegio, llueve xk son vacaciones y en vacaciones siempre llueve yo creo que es por joder!

which says - it is not raining because you got out of school, it rains because these are holidays ...

Another case in which we observe a tendency to alter the spellings of the word "because" in a sentence is when the sentence has a nested structure. For example

- April 4th, 2012: I isolate myself when I feel a certain type of way & I'm trying to get it off my mind because I don't like for people to be down becuz of me

In Spanish, the word "porque" means both "because" as well as "why". Observing all occurrences of both in the week of 26 March - 4 April 2012, 8 out of 9 used "xk" for why and "porque" for "because", and only one case was the opposit. This suggests that new spellings can be used to transfer more information on words that looked the same in previous spellings.

## VIII. COMPARISONS WITH OTHER TYPES OF CONTENTS AND TOOLS

In this Section we compare the frequency of occurrence of "because" in its different spellings obtained in Twitter to the frequency of its appearance in the World Wide Web. For the latter we used "google search". We repeated the same

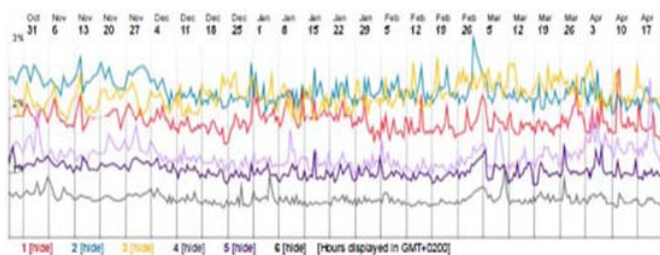


Fig. 18. The relative frequencies of the integers 1, 2, 3, 4, 5, 6

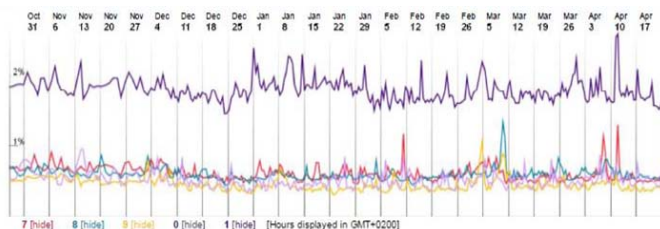


Fig. 19. The relative frequencies of the integers 7, 8, 9, 0, 1

experiment restricting to those pages that fall into the category of "news" under google.

Our findings are summarized in the Table II.

The table presents the normalized popularity: at each row, we divided the corresponding number by the first one in that row. In that way we can compare the relative "popularity" of each form of "because" on different media.

The fractions of the versions of "because" that are obtained by shortening, "becuz" and "bcuz", are seen to be much higher in twitter than those obtained in google and in google news. The number of appearances of each one of these two versions, divided by the number of times that "because" appears, is more than 500 times larger in twitter than over the whole Internet. It is more than 100 times larger in twitter than it is over news documents found by twitter.

This ratio in twitter is also larger than in the whole internet for the versions obtained by further clapping the word "because" (where the first syllable disappears) with one exception: the word "cause" which is used more on the Internet since the other meanings of the word "cause" appear more frequently there. This is also true to the finding over google news.

Two other exceptions occur with respect to google News: "cz" is very frequent there as it is used with the meaning of the Czech Republic. The spelling "cos" also has many other uses that appear in google News.

## IX. 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.

[RyanElliott](http://t.co/1BQFm7N3) : <http://t.co/1BQFm7N3> 3 more fans til 100 like it guy, pleeeeeeease <3

on June 5 at 05:58 p.m.

[themightybbouch](#): A dream you don't fight for can haunt you for the rest of your life <3

on June 5 at 05:58 p.m.

Fig. 20. Examples of tweets using "3"

- 3 is used frequently combined with < which gives <3. This can be interpreted as a kiss or a heart, see Fig 20.

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.

For more on the role of numbers in twitter, see [3].

## X. CONCLUSIONS

This paper contributes to identifying and understanding some of transformations in spelling and use of words over twitter. This includes a geo-linguistic analysis that allows one to track different type of transformations in different communities that have a common language in common. Among the many examples presented here, we have studied in more detailed the transformations of the word "because", both in English as well as in Spanish. We saw that some common form of using two different spellings in the same sentence has emmerged both in Twinglish as well as in twitter-Spanish. We managed to differentiate between the locations of various versions of "because".

The creation of new spelling and forms of words in twitter 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 comfortable as that of a laptop.

We saw that in the creation process of new spellings, alpha-numerical symbols often replace cylables according to

Spelling:	because	bcuz	becuz	cause	cuz	cos	cus	coz	cZ
Normalized popularity									
trendistics	100	2.00	1.41	61.3	22.7	5.94	3.69	1.64	1.55
Google	100	0.00251538	0.00264231	26.8461538	0.8076923	1.0384615	0.0869231	0.2115385	0.9769231
Google News	100	0.0171975	0.0140127	122.2929936	0.4356688	11.1464968	0.656051	7.1974522	48.4713376

TABLE II  
NORMALIZED POPULARITY OF VARIANTS OF BECAUSE IN %

- (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 pronunciation 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 tolerance 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 Twitternauts do not seem eager to drop them completely (as we saw in the shortning of "because").

We saw that Twitternauts often convey the accents they use. This was the case of the word "because" whose twitter 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 pronounced as a "d". We illustrate this in Figure 21.

#### Acknowledgement

The authors wish to thank C. E. Contreras Oller (Department of modern languages, Univ de Los Andes, Mérida, Venezuela), Prof Aharon G. Kleinberger (Department of Arabic Language and Literature, Haifa University, Israel) and Julien Gaillard (Univ of Avignon).

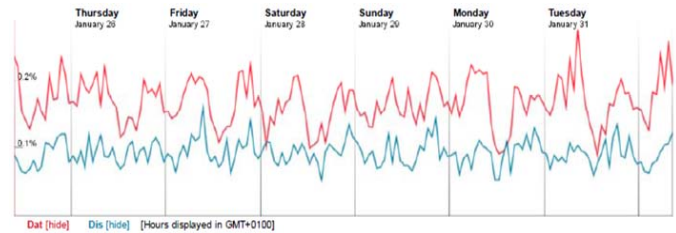


Fig. 21. The words "this" and "that" spelled as "dis" and "dat"

#### REFERENCES

- [1] O'Connor, B.; Balasubramanian, 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).
- [3] Stephan Gouws, Donald Metzler, Congxing Cai and Eduard Hovy, "Contextual Bearing on Linguistic Variation in Social Media", Proceedings of the Workshop on Language in Social Media, June 23, 2011.
- [4] Hong-mei Sun, "A Study of the Features of Internet English from the Linguistic Perspective", Studies in Literature and Language, Vol. 1, No. 7, 2010, pp. 98-103.
- [5] Isabel Gretel María Eres Fernández and Paulo Augusto Almeida Seemann, "A study on language changes of written spanish in internet chats" Trab. linguist. apl. vol.48 no.1 Campinas Jan./June 2009
- [6] Zoe Rimay, "Cybercultural Communication", Budapest Univ. of Technology and Economics, Faculty of Economic and Social Sciences. Report.