Texting in Newsgroups

How technology may influence a language

Marco Stefano Tomatis
dept. of Cultures, Politics and Society
University of Turin
Turin
marcostefano.tomatis@unito.it

Abstract—**This paper deals with a particular phenomenon which is strictly tied with modern communication technology; the usage in newsgroup messages of those particular abbreviations that one can usually read in mobile phone messages. The texting abbreviation means the orthographic substitution of entire standard words with a completely different set of graphemes according to phonetic patterns. In order to elaborate this work, a corpus-based enquiry has been made and a number of analysis which took into account the general topic of the message the abbreviation occurred in, were carried out.**

Keywords- Textings; corpus linguistics; newsgroups; orthography; phonology

I. INTRODUCTION

This paper aims at investigating how English orthography may be used in a creative way, in particular in computer-mediated communication (CMC) contexts. In many cases, like when exchanging short messages with mobile phone systems, also known as "texting", the need of reducing the average amount of characters per word has produced different sort of abbreviations[1]. This process is, in general, very productive in informal writing and can be found in different languages according to both the specific word which has to undergo the reduction process and the orthographic rules of the language involved. Differently from other researches which tend to analyze in a comprehensive way the whole phenomenology of texting, describing typical acronyms and emoticons, in the present study I will face the phonetic-driven orthographic shortenings only. In particular, I will try and demonstrate that the English words which are entitled to undergo the orthographic remodelling process are affected by a variable productivity rate in accordance with the specific topic of discussion.

[1] As historical information, it is interesting to notice that in spite textual abbreviations were massively used in mobile phone short text messages, their first usage can be found in Information Technology, in particular in early 1990s MS-DOS operating system. Because of technological limitations affecting its file system, the maximum length of MS-DOS filenames was limited to eight characters. Therefore, in order to avoid such an annoying limitation, programmers were induced to use an alternative spelling to express the same meaning with a reduced number of characters. Program names like "dos2unix" are a clear example of the said abbreviation technique.

II. THE ORTOGRAPHIC ABBREVIATION ISSUE

A. *Technology and written communication*

Computer-mediated communication is defined as any form of linguistic text-based interaction occurring between two or more users of networked computers or electronic devices. Therefore, since this way of communication uses written texts only, it is strictly tied to orthographic rules. However, differently from other languages like Italian or German, where orthography is quite clear and straightforward in terms of correspondence between the graphematic and the phonological layer, English retains and maintains historical orthographic traditions, causing the pronunciation of standard English to change significantly from its normatively regular spelling. As a consequence of this, in particular situations (i.e. when the author of a message is forced to respect a rigid limitation in the number of charcters that he can use, as in texting) such a peculiarity may easily lead to the creation of semantically well-formed new words made just of one or two characters, which are totally homophonous to the corresponding standard form when pronounced as a whole single element.

B. *A corpus-driven approach*

The present study was produced by analyzing the UK subset of the wider corpus called "NUNC"[2] which was developed at the University of Turin by collecting text messages from a number of Usenet newsgroups. In order to create a balanced, multilingual, parallel corpus of contemporary language, the entire amount of data was firstly divided into five different subcorpora according to the different languages it was made of (i.e. Italian, English, French, Spanish, German). Secondly, each language specific subset was split into a number of subject-driven subcorpora to help users carry out multilingual researches on different topics. As regards this paper, it will focus on the NUNC UK corpus only, to investigate the most interesting cases of orthographic reformulation and describe the context they were found in. As confirmed by previous researches in this area, we noticed that only a very specific set of both alphabetic and numeric characters are commonly used, which are limited to seven elements only. Yet, it is important to keep in mind that only standard characters with a specific phonological feature were

[2] The different, multilingual subsets of corpora which constitute the whole NUNC corpus are freely accessible at the web address: http://www.morfoweb.it/bmanuelorg/projects/ng-HOME.html

taken into account. Other elements like the symbol "@", which in computer science is conventionally used to represent the preposition "at", were left aside on purpose. The following table shows the correspondence between the characters used, their phonologic value and their orthographic counterpart

TABLE I.

Phonetic-Orthographic Corresponsion		
Character	***Phonetic Transcription***	***Meaning***
B	bi:	Be
C	si:	See
R	ɑ:	Are
U	ju:	You
Y	waɪ	Why
2	tu:	Two
4	fɔ:	For

Regarding the orthographic aspect, Crystal [3] states: "the use of single letters, numerals, and typographic symbols to represent words […] are technically known as logograms or logographs. […] Logograms in texting may be used alone, or in combination". Therefore: "It is the pronunciation of the logogram which is the critical thing, not the visual shape." This statement implies that the above listed characters may be divided into two groups according to their own linguistic nature. Indeed such graphemes may either behave as free morphemes (e.g. "C U" = see you), or get combined with other elements to produce an entire word (e.g. "B4" = before).

III. DATA ANALYSIS - AN OVERVIEW

A. Abbreviations and Acronyms

The results of the quantitative analysis calculated by taking into account the complete NUNC UK corpus, prove that the usage of the grapheme "U" is the most relevant, scoring 318 occurrences. After a significant gap we found the letter "R", which scores a total amount of 28 occurrences only. Then, the numeric characters "2" and "4" follow, showing 14 and 10 occurrences respectively. The lowest rankings are covered by the elements "B" , "C" and "Y", tied to 7, 2 and 1 occurrences. As regards the words created by combining two different elements, in the whole corpus only the forms "UR" (= your), "B4" (= before) and "NU" (= new) were found, assuming a score of 42, 25 and 4 occurrences.

TABLE II.

Usage of Texting Abbreviations		
Character	***Free morpheme***	***Part of a word***
B	+	+
C	+	-
R	+	+
U	+	+
Y	+	-
2	+	-
4	+	+

The reasons for the particular pattern which comes out from the table above may be due to a number of reasons, ranging from the stylistic choices of the authors to the grammar function the described elements cover. Regarding this specific aspect, data clearly show that the element used as a free morpheme representing the pronoun "you", reached the highest rate. Yet, when used in combination with other elements to create a whole new word, its ranking drops impressionably down. To enforce the hypothesis that the grammar function represents an important factor to change the occurrence values, it is worth to notice that none of the other elements substitute a pronoun; they may only represent verbs ("B", "C" and "R"), prepositions ("2" and "4") and adverbs ("Y"). As regards this last character, it is interesting to notice that besides assuming the meaning of "why", in the NUNC corpus the grapheme "Y" is more commonly found as an abbreviation of the pronoun "you". The data taken from the corpus, however, showed us that in such case the letter "Y" is always followed by an apostrophe sign (e.g. "y'all are ridiculous"). This particular behaviour appears to be functional for establishing an effective communication; it helps the reader of the message disambiguate between the interpretation of the character as a texting shortening and its usage as a mere abbreviation of the standard form. Naturally, in accordance with the aims of this paper, the letter "Y" will not be taken into account when used as a simple reduction without involving any phonologic aspects. Although the alphabetic elements do not create particular problems from a statistic point of view, we cannot affirm the same for the numeric values. Indeed, without adopting an accurate, human driven Part-Of-Speech tagging, their presence in the text raises a number of problematic issues, starting from the creation of a simple frequency list.

B. Statistical distribution

A more interesting analysis regards the numeric distribution of the characters used as a texting abbreviation within the different subcorpora which are part of the whole NUNC UK corpus. The results of our study show that the largest usage of the said abbreviations have occurred in the corpus that includes all the newsgroups dedicated to the world of motors. In such corpus, the total amount of the shortening patterns examined is equal to 344 occurrences, most of which are represented by the letter "U" (250 occurrences). Like the figures regarding the frequency distribution of the alphanumeric elements, even in this case an interesting big gap appears. Indeed, the second ranking position is taken by the corpus containing discussions about business issues. Collected data show that in this corpus only 76 elements related to texting were used, while less than a half of the previous occurrences appeared in the corpus related to cooking (26 occurrences). At the bottom of our list is the

corpus hosting discussions about photography, where only 5 texting abbreviations were found. Although the above figures may appear a little idiosyncratic at a first sight, a more specific analysis which takes into account both the message topics and the graphemes one can find in the different corpora we took into account, reveals useful. Although in general the language of the messages posted in public access electronic boards can be considered sloppy, careless of grammar rules and, in some ways, rude, in the newsgroups dealing with motor issues the characteristics pointed out before seem to be more widespread. Sentences like: "Which 1 of them 3 wud b best ???", "so y is the cameras there ?" and "just a little troll 2 c if u idiots were as stupid as u seem" are clear examples of the everyday language style which is adopted in discussions about cars and motors. In the "motors" corpus, this stylistic aspect may be a consequence of the stereotype of the uneducated, grossly mannered truck drivers or motor mechanics. Quarrels between newsgroup users raise frequently, in particular when they use bully manners to prove their superiority in car knowledge and driving ability. The following is an example of a such an exchange:

"if u want 2 c my car then come down 2 southampton and i will take u 4 a drive.....but deffo not posting pics 2 pander 2 anoraks goll

> > > > > > open invite 2 come down POWERHOUSE and see exactly what my 350bhp rs is going 2 do your 405, xl125, saab turdo.....whatever!

> > > > goll

> > > > POWERHOUSE BOYZ RULE!

> > > Go bang and throw smoke through my windows/helmet?

> > > Plus - southhampton is on the south coast yeah? It would take me several days to get there on a 125...

> > > --

> > > Dan

> > trouble with cyber anoraks dan is that they can never back up the nonsense

> > they post here!

> Except that they can.

> The only person that posts random statements, is YOU !

> If you've got a 350bhp RS, then post some pics and show the world. Give us info, and we'll take you seriously. Until then, you're just a troll."

Right on the opposite is the corpus made of discussions about photography, where the language style is definitely more solid, firm and respectful of the standard English rules. Again, like before, this aspect is tightly related to the particular argument of discussion. Photography, on the one hand, is generally considered to be an art like sculpture or painting, so people who are interested in such a subject have a relatively high education level. On the other hand, photography is a very complex matter; it leads the newsgroup users to ask for help or discuss about technical issues using very specific, appropriate terminology. As a support to this statement, we may observe that the corpus which contains newsgroup postings about photography shows a total amount of 5 texting patterns only, all of them used as a substitution of the pronoun "you". In the middle between the opposite poles above lay the corpora about business and cooking. As already mentioned before, if compared with the corpus about motors, the corpus about business shows a reduced number of such abbreviations, which percentage is only around the 22%. This low usage of non-standard forms in newsgroups related to business may be explained adopting the same considerations we made about photography. Also in this case the particular subject forces users to adopt a formal, technical language, avoiding uneducated style of writing as far as possible. Yet the business corpus is ranked in second position, just before cooking. After the discussion above, one would expect a different distribution of texting patterns throughout the corpora which take into account business and cooking newsgroups. The reason for this apparently odd results lies again in the peculiarities of the newsgroup related to business. Indeed, notwithstanding the formal seriousness of this subject, which reflects in the style of the language adopted, the newsgroups about business tend to attract a number of users who post commercial messages claiming to be able to provide easy financial gains. Those messages are designed to be communicatively effective and convince as more people as possible, consequently they tend to induce curiosity in the reader adopting an everyday language style. An example of such a practice may be found in the following post:

"Hi My Name is Monica. Do Check out My Page with my Pics and All about me.You sure will love it. How I Earned 50,000 $$ From Web With 0 $$ INvestment I don't think one need investment to earn money one only need his/her mind to be focused to earn money. Their is a lot of money to be made one internet. People are making tons of it as well.If u wanna learn i have share a lot of my secrets on my FREE site ofcouse tho i have few dot coms but i prefer free sites coz its easy for others to make one as well. CHeck it out."

Summing up, the elements which have been taken into account follow inside the four corpora the distribution below:

TABLE III.

Distribution of Texting Abbreviation					
	Cooking	**Motors**	**Photo**	**Business**	**Total**
B	0	4	0	3	7
C	0	2	0	0	2
R	1	23	0	4	28
U	20	250	5	43	318
Y	0	1	0	0	1
2	1	3	0	10	14
4	2	1	0	7	10
UR	1	35	0	6	42
NU	0	4	0	0	4
B4	1	21	0	3	25

Distribution of Texting Abbreviation					
Total	26	344	5	76	

IV. USAGE EXAMPLES

The following examples describe the linguistic use of the characters listed above, specifying the number of occurrence found in the different subcorpora of the NUNC UK and their specific operative context. All the results have been semi-automatically filtered to eliminate the semantic ambiguity and the repeated text patterns. All the examples below have been extracted from the UK corpora, so they may show differences if compared with the original messages because of tokenization and other text processing actions. Due to space limits, for each case the reported examples have been limited to the most significant entry.

UK_cooking

R: 1 occurrence

“turning everything into a plastic world of Crap Pints R Us”

U: 20 occurrences

“Catch u all soon I hope . PS any lurkers , it would be good to hear from u , whatever u had to say .”

UR: 1 occurrence

“pile up a 14 gm double shot to ur not necessarily expensive espresso machine portafilter”

B4: 1 occurrence

“Just don't go down the pub b4 u shower and change !”

2: 1 occurrence

“calling 2 u on 0800 083 0501”

4: 2 occurrences

“Thought I'd take a look at Spice4u's buffet .”

UK_motors

B: 4 occurrence

“your car should b in tip top condition”

C: 2 occurrences

“if u want 2 c my car”

R: 23 occurrences

“they r only used for slippy times as such mud etc”

U: 250 occurrences

“I will let u know what I thought of my first grand prix experience !”

Y: 1 occurrence

“so y is the cameras there ?”

UR: 35 occurrences

“look after ur car”

NU: 4 occurrences

“it seems nu venture owners club”

B4: 21 occurrences

“so i need advice of people who have done it b4”

2: 3 occurrences

“come down 2 southampton”

4: 1 occurrence

“i will take u 4 a drive”

UK_photo: 5 results

U: 5 occurrences

“If u want tell me what u think about my photos .”

UK_business: 37 results

B: 3 occurrences

“not 2 b confused with a trading plan”

R: 4 occurrences

“Thought : no wonder , you r sage reseller”

U: 43 occurrences

“why are u insisting on the fact that i 'm a charlatan”

UR: 6 occurrence

“ok that 's ur business”

B4: 3 occurrences

“Anyone intrested b4 i sell it on eBay”

2: 10 occurrence

“I 'll deliver 2 to you for £5”

4: 7 occurrences

“im looking 4 a business plan 4 trading”

REFERENCES

[1] Y. Park, R. J. Byrd., “Hybrid Text Mining for Finding Abbreviations and their Definitions”, in Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMPL ’01). Washington D.C.: Morgan Kauffmann , 2001, pp. 126-133.

[2] E. Pistolesi, “Il parlar spedito. L’italiano di chat, e-mail e sms”. Padova: Esedra, 2004.

[3] D. Crystal, “Txtng: the gr8 db8”. Oxford: Oxford University Press, 2008.

[4] G. Myers, “The Discourse of Blogs and Wikis”. London: Continuum, 2010.

[5] C. Tagg, “Wot did he say or could u not c him 4 dust? Written and Spoke Creativity in Text Messaging”, in C. Ho, K. Anderson and A.

Leong (eds) Transforming literacies and Language. London: Continuum, 2011, pp. 223-236.

[6] C. Tagg, "Discourse of Text Messaging: Analysis of SMS ommunication", London: Bloomsbury, 2012