

# ¿k dizez? A corpus study of Spanish Internet orthography

Mark Myslín and Stefan Th. Gries

University of California, Santa Barbara, CA, USA

## Abstract

New technologies have always influenced communication, by adding new ways of communication to the existing ones and/or changing the ways in which existing forms of communication are utilized. This is particularly obvious in the way in which computer-mediated communication (CMC) has had an impact on communication. In this exploratory article, we are concerned with some characteristics of a newly evolving form of Spanish Internet orthography that differ from standard Spanish spelling. Three types of deviations from 'the norm' are considered: a reduction (post-vocalic *d*/[ð] deletion in *-ado*), a transformation (namely the spelling change from *ch* to *x*), and reduplication (of characters). Based on a corpus of approximately 2.7 million words of regionally balanced informal internet Spanish compiled in 2008, we describe the spelling changes and discuss a variety of sometimes interacting factors governing the rates of spelling variants such as overall frequency effects, functional (pragmatic, sociolinguistic, and iconicity-related) characteristics, and phonological constraints. We also compare our findings to data from Mark Davies's (2002) Corpus del Español (100 million words, 1200s–1900s, <http://www.corpusdelespanol.org>) as well as other sources and relate them to the discussion of the register/genre of Internet language.

### Correspondence:

Stefan Th. Gries  
Department of Linguistics,  
University of California  
Santa Barbara,  
Santa Barbara,  
CA 93106-3100, USA  
E-mail:  
[stgries@linguistics.ucsb.edu](mailto:stgries@linguistics.ucsb.edu)

## 1 Introduction

New technologies have always influenced communication, by adding new ways of communication to the existing ones and/or changing the ways in which existing forms of communication are utilized. This is particularly obvious in the way in which computer-mediated communication (CMC) has had an impact on communication. One very obviously visible way in which CMC has been influencing communication is the large number of new linguistic expressions such as 'regular' words (e.g. *bcc*, *blog*, *podcasting*, etc.), emoticons and similar symbols (e.g. ':-)', ':-|', ':-S', etc.), abbreviations standing for complete phrases (e.g. *lol* for 'laughing out loud', *brb* for 'be right back', IMHO for 'in my humble opinion', *AFAIK* for 'as far as I know', etc.).

In this article, we are concerned with an aspect of communication that is often regarded as somewhat peripheral, namely orthography. CMC and other forms of electronic discourse have given rise to forms of orthography that deviate from standardized conventions and are motivated by segmental phonology, discourse pragmatics, and other exigencies of the channel (e.g. the fact that typed text does not straightforwardly exhibit prosody). More specifically, we will explore several new trends in the orthography of Internet Spanish, which is by now the third most widely used language on the Internet (Fig. 1).

In keeping with the dominant role of English on the Internet, there is now quite a lot of work on Internet English. However, in spite of its growing importance, there is still very little work on the

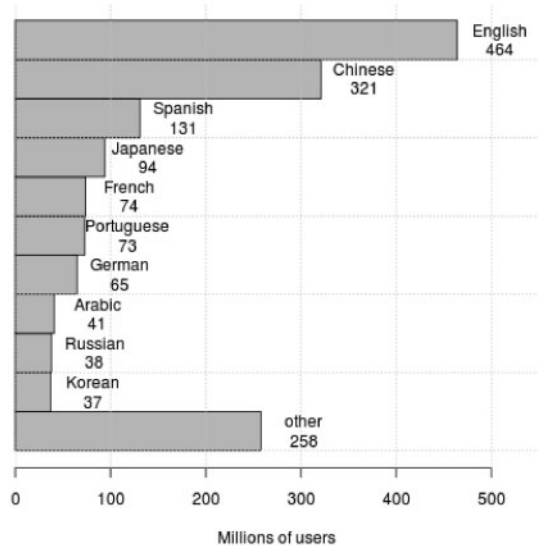


Fig. 1 Top 10 languages on the Internet (in millions of users; Internet World Stats 2009)

characteristics of Internet Spanish (e.g. Cervera 2001, Morala 2001, Moreno de los Rios 2001, and Llisterri 2002). Most of these studies are strictly impressionistic, offer no quantitative data, and address only the chat genre, some under the assumption that it is representative of all Internet Spanish. In fact, there is not an even modestly comprehensive overview of the many different facets of Internet Spanish, which, when combined, can change the orthographic characteristics of standard Spanish considerably. Cf. (1) for an example of Spanish Internet orthography (hereafter SIO) with its standardized orthography in (2).

- (1) hace muxo k no pasaba x aki,, jaja,, pz aprovechio pa saludart i dejar un komentario aki n tu space q sta xidillo:)) ps ia m voi <<http://profile.myspace.com/index.cfm?fuseaction=user.viewprofile&friendID=198138943>>
- (2) Hace mucho que no pasaba por aquí, jaja. Pues aprovecho para saludarte y dejar un comentario aquí en tu space que está chidillo. Pues ya me voy.

SIO cannot be characterized as a rigid one-to-one grapheme mapping from standard Spanish since, while being somewhat systematic in some respects, it also exhibits considerable internal variation. For example, in (1) above, *que* is spelt in two different ways: *k* and *q*. In this article, we attempt to explore and characterize several of the most visible ways in which SIO differs from standard Spanish. Therefore, before we discuss a few case studies in more detail, we would like to provide a brief overview of the kinds of patterns we observed in our corpus (whose makeup will be outlined below) for future work on this topic. We classified the deviations from standard Spanish spellings into two categories: one with differences that were fairly clearly related to informal Spanish phonology, and one where phonological relations were much less apparent. The distinction between the categories was done heuristically; nothing theoretically relevant hinges on it (cf. Tables 1 and 2 for overviews).<sup>1</sup>

Obviously, space does not permit a full-fledged analysis of all these ways in which SIO differs from standard Spanish orthography. In this largely exploratory article, we therefore decided to focus on three different mechanisms by which SIO differs from standard Spanish:

- a deletion, namely from *-ado* to *-ao*;
- a change, namely from *ch* to *x*;
- repetitions, e.g. from *hola* to *hoola*.

The remainder of this article is structured as follows. Section 2 discusses our data and methods, in particular how we compiled a corpus of SIO. Sections 3, 4, and 5 discuss our case studies in detail, providing detail on the retrieval and cleaning of the data as well as the quantitative methods we used, the linguistic factors we studied, and the results. Section 6 concludes. One terminological remark is in order: although the medium of communication is, strictly speaking, written, we will refer to interlocutors and their communication as *speakers* and *utterances* because SIO, while shaped by the medium, exhibits many of the characteristics of spoken language (cf., e.g. Baron 2000, 2003 and Crystal 2001 for good overviews of the different kinds of CMC and some of their characteristics).

**Table 1** Phonologically motivated features of SIO

Standard orthography	Internet orthography	Examples	Phonological correlate
([aeiou])[bdg]([aeiou])	\12	<i>hablabas</i> → <i>hablaas</i> <i>saludos</i> → <i>saluos</i> <i>me gusta</i> → <i>me usta</i>	Intervocalic voiced plosive elision
u([aeio])	w\1	<i>buena</i> → <i>wena</i> <i>igual</i> → <i>iwat</i> <sup>13</sup>	Pre-/w/ voiced plosive elision
([aeiou])s	\1h	<i>somos</i> → <i>somoh</i>	Post-vocalic /s/ debuccalization
([aeiou])s	\1	<i>llegamos</i> → <i>llegamo</i>	Post-vocalic /s/ elision
^es[^aeiou]	^s?	<i>espero</i> → <i>spero</i> <i>está</i> → <i>ta</i>	Pre-/sC/ /e/ aphaeresis (can combine with post-vocalic /s/ elision)
([aeiou])[bv]	\1v	<i>iba</i> → <i>iva</i>	Post-vocalic /b/ spirantization
([^aeiou])[bv]	\1b	<i>vemos</i> → <i>bemos</i>	Non-post-vocalic /b/ is a plosive
h		<i>hacer</i> → <i>acer</i>	<i>h</i> has no phonetic value
ch	sh	<i>echo</i> → <i>esho</i>	/tʃ/ deaffrication

**Table 2** Non-phonologically motivated features of SIO

Standard orthography	Internet orthography	Examples
[sz] c([ei])	[csz] [sxz]\1	<i>hermosa</i> → <i>hermoza</i> <i>hice</i> → <i>hize</i> <i>hizo</i> → <i>hiso</i> <i>hace</i> → <i>haxe</i>
[usz] ch	x	<i>cuidate</i> → <i>cxidate</i> <i>hizo</i> → <i>hixo</i> <i>mucho</i> → <i>muxo</i>
t([aeiou])	th\1	<i>besitos</i> → <i>besithos</i>
c([aou]) qu([ei]) g([^ei])	k\1	<i>poco</i> → <i>poko</i> <i>cuidate</i> → <i>kuidate</i> <i>quiero</i> → <i>kiero</i> <i>agrega</i> → <i>akreka</i>
cu	qu	<i>cuando</i> → <i>quando</i>
[iy] ll	[iy]	<i>muy</i> → <i>mui</i> <i>mis</i> → <i>mys</i> <i>llego</i> → <i>iego</i> <i>llamar</i> → <i>yamar</i>
([dmtq])u?e\$	\1	<i>porque</i> → <i>porq</i> <i>te</i> → <i>t</i>
ie	e	<i>quiero</i> → <i>kero</i>

## 2 Data

As a first step, we needed to compile a corpus of informal SIO. To that end, we used the scripting language R to crawl selected forums and social networking web sites (cf. Gries 2009 for details as well as R Development Core Team 2008). In May 2008, we compiled a corpus of approximately 2.7 million words of informal Internet Spanish, consisting of user-generated descriptions of photos and videos, as well as comments on these and postings on social networking site profiles (which, although generally termed *comments*, often express greetings and

messages rather than stance toward the actual profile pages). The mean length of entry in the corpus is 19.5 words ( $sd = 36.2$ ). Table 3 provides an overview of the web sites from which the data were obtained.

While it is hard to assess to what degree this corpus is representative of, or balanced with regard to, Internet Spanish, we consider it relatively representative in the sense that the highly personal discourse of the social networking sites and the less intimate, more diversified discussions of the photo and video sites should go some way to represent differently involved sub-categories of

**Table 3** Composition of the Spanish Internet Orthography corpus

Website	Genre	Approximate percentage of corpus
www.fotolog.com	Comments	43
www.hi5.com	Comments	27
www.fotolog.com and www.youtube.com	Descriptions	21
www.youtube.com	Viewer comments	9

Internet language. In addition, further efforts were made to ensure some degree of dialectal representativity, as Spanish varies widely by country and region. To this end, we used the search-by-country feature of both Fotolog and hi5.com and selected the first three users from each official Spanish-speaking country. Using the friend lists of each of these three users, the R scripts indiscriminately harvested all of the comments on the profile pages of each of these friends (each of the three country representatives had between 100 and 200 friends). A surprising majority of the country representatives' friends were in fact from other Spanish-speaking countries, which seemed roughly distributed by population, with Mexico, Spain, Argentina, and the USA well represented. No measure was taken to 'correct' this phenomenon, as it is a kind of self-balancing middle ground between equal representation of different geographic varieties of Spanish and proportional representation based on numbers of speakers of each variety.<sup>2</sup> No regional sorting feature exists for YouTube videos, so the sampling method was simply to use the web site's search-by-language function and then use R to automatically harvest all comments and descriptions for several thousand of the most viewed videos uploaded by Spanish-speaking users.

In order to compare our SIO data to other data, we utilized two other sources. First, we used Mark Davies's (2002) 100 million word Corpus del Español (CdE) as a reference corpus to represent standard Spanish orthography. Second, since much Internet discourse involves many colloquial and vulgar terms, we also compiled a list of general Spanish vulgarities in all of their inflections based on the list of vulgar-tagged words on the Wiktionary open-source Spanish dictionary, since this Internet-user-generated list seemed more inclusive and up-to-date than formally

published dictionaries (<<http://en.wiktionary.org/wiki/Category:es:Vulgarities>>, accessed June 1, 2009).

### 3 Reductions in Spelling: Post-vocalic [ð] Deletion in Words Ending in *-ado*

#### 3.1 Introduction

The first feature of SIO we investigate is the deletion of a single character in a way that reflects pronunciation in certain speech varieties. Intervocalic voiced stops are generally spirantized but can be deleted completely in the onset of an unstressed syllable in rapid or informal speech, with *d* being the most commonly affected segment (cf., for example, Piñeros 2009, p. 319). Llisterri (2002, p. 69) reports *d* as the most commonly elided word-interior segment in his chat corpus, and looks closely at words ending in *-ado*, generally a past participle marker, and its various inflections (pp. 73–76), correlating orthographic omission with colloquial and especially Andalusian Spanish phonology.

#### 3.2 Methods

In our own investigation of *d*-deletion, we focus only on *-ado* without its feminine and plural-inflected variants *-ada*, *-ados*, and *-adas*. In order to compare frequencies of words with deletion to words without deletion, we searched our corpus for all words ending in *-ado* or *-ao*. To avoid interference from phenomena other than *d*-deletion, such as apparently typographically erroneous *d*-insertion, we immediately discarded the handful of word forms that end in *-ao* in standard Spanish spelling, such as *cacao* 'cacao'. We took these words to be those occurring in the CdE more than five times

**Table 4** Type frequencies of *-ado/-ao* forms in the SIO corpus and in Llisterri (2002)

	SIO corpus	Llisterri's (2002) chat corpus
only <i>-ado</i>	893	96
only <i>-ao</i>	419	104
<i>-ado</i> and <i>-ao</i>	277	65
Total	1589	265

**Table 5** Distribution of the two spelling variants across both corpora

Data	<i>-ado</i>	<i>-ao</i>	Total
SIO corpus	571	91	662
CdE	6423	2	6425
Total	6994	93	7087

with *-ao* but never with *-ado*. We further refined our list of *-ado* and *-ao* matches by discarding

- alternate spellings of the above-mentioned standard *-ao* words, such as *shao* in the case of *chao* ‘ciao’;
- the proper name *Pao*;
- *nao* when occurring as the Portuguese *não* ‘no’;
- standalone occurrences of *ado* and *ao*.

Finally, since English words are not infrequent in the corpus,<sup>3</sup> we checked for English words by comparing our list of matches with words occurring in the British National Corpus over five times (based on Kilgarriff 1996), but did not find any matches that were not also Spanish words. For this study, we then decided to examine the 50 most frequent forms (by combined *-ado* and *-ao* occurrence).

### 3.3 Results 1: our SIO corpus versus Llisterri's (2002) chat corpus

With the above considerations, we found 1589 types of *-ado/-ao* words that were distributed as shown in Table 4 (with Listerri's (2002) type frequencies for comparison).<sup>4</sup>

To determine whether the frequencies of types that allow *-ao* differs between Llisterri's data and ours, we computed a chi-square test for independence on the italicized bottom two rows of Table 4. According to this test, the two data sets do not differ

significantly with regard to the numbers of forms that take *-ado* and *-ao* versus those that only take *-ao* ( $\chi^2=0.1$ ,  $df=1$ ,  $P\approx 0.75$ ).<sup>5</sup>

One conclusion from this is that, while the two genres differ in terms of interactivity—more interactive chat data of Llisterri (2002) versus less interactive comment/description data in our SIO corpus—they exhibit the same degree of *d*-deletion so characteristic of informal speech.

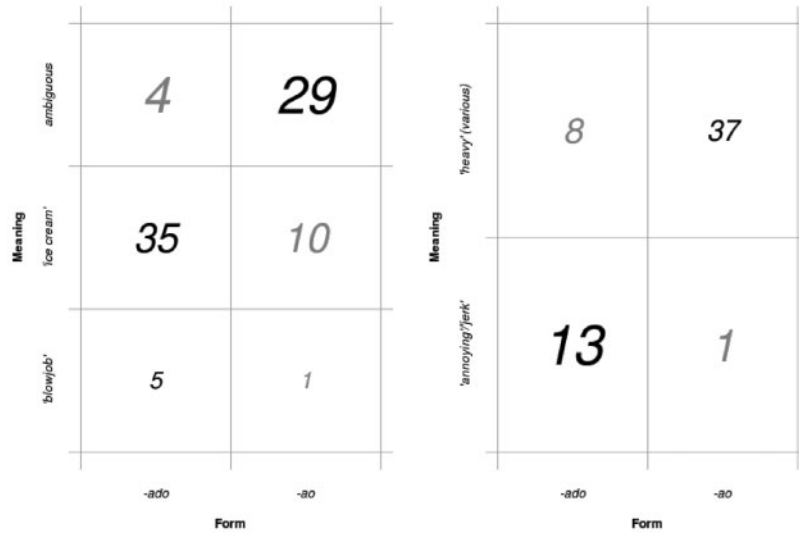
### 3.4 Results 2: the 50 most frequent words in the SIO corpus versus the CdE

In order to examine *d*-deletion in *-ado* more closely in frequent words and quantify differences with non-Internet Spanish, we compared the spellings of the fifty most frequent forms in our list (after the above-mentioned modifications) with their spellings in the CdE.<sup>6</sup> For each word type we constructed a 2x2 table of the kind exemplified in Table 5 (on the basis of the word *pasado* ‘past, passed’). For this kind of table, we then determined the percentage of *d*-deletion in each corpus for all word forms (for *pasado*, 91/662 or 13.75% in the SIO corpus, and 2/6425 or 0.03% in the CdE). Figure 2 shows the difference in percentage of *d*-deletion in the SIO corpus and the CdE as a function of overall frequency in the SIO corpus (both axes are on a logarithmic scale). Plotted word forms represent the word's percentage of *d*-deletion versus overall frequency in SIO, and corresponding diamonds linked by dashed lines represent the word's percentage of *d*-deletion in the CdE (when the word was attested in the CdE).

In Figure 2, *d*-deletion is far more frequent in our Internet corpus than in the CdE. Most of the 50 most frequent word forms occurred with deletion zero times in the CdE, and only a handful of these did the same in SIO. However, *d*-deletion does not simply apply across the board; rather, there are several, sometimes competing or interacting factors that motivate different proportions of *d*-deletion.

First, there is the factor of word frequency. In SIO, the percentage of *d*-deletion appears to have a roughly inverse relationship to frequency, so that more frequent words tend to exhibit less deletion. One reason for this may be that the most frequent words are more entrenched in the speakers'





**Fig. 3** The interaction of meaning and spelling for *helado* (left panel,  $\chi^2 = 35.38$ ;  $df = 2$ ;  $P < 0.001$ ;  $V = 0.65$ ) and *pesado* (right panel,  $\chi^2 = 26.26$ ;  $df = 1$ ;  $P < 0.001$ ;  $V = 0.68$ )

construct a distinct style/social identity in the way their spelling reflects two interrelated rules: ‘modify words that have special pragmatic functions and, if you are really determined to modify a common-or-garden kind of word, then make big/several changes.’<sup>7</sup>

A final determinant is phonology. The only two words in the top 50 that are not stressed on the penultimate syllable and thus virtually never undergo *d*-deletion in speech, *sábado* ‘Saturday’ and *agradó* ‘pleased’, do not exhibit *d*-deletion at all in SIO.

### 3.5 Results 3: vulgar in SIO

We have already seen that *d*-deletion is more frequent among words with a special pragmatic function. This is confirmed by a closer look at both the vulgar terms represented in Figure 2 and a comparison with the words listed as vulgar in our Wiktionary source.

#### 3.5.1 Vulgar words among the 50 most frequent words

Among the top 50 *d*-deleted words in the corpus, slang and vulgar words exhibit the highest proportion of *d*-deletion. In Figure 2, the six forms with the highest percentage *d*-deletion, which are

relatively clearly differentiated from the bulk of the data, fall into this category: three words that are attested in both the SIO corpus and the CdE (*agado* ‘fucked up’, *pesado* ‘heavy, annoying, jerk’, and *helado* ‘ice cream, blowjob’) and three that are only attested in the SIO corpus (*qliado/culiado* ‘motherfucker’, *aweonado* ‘asshole’ (standard spelling *ahuevonado*), and *pelado* ‘thug, dude’). *Qliado/culiado* and *aweonado* are not attested in the CdE with or without deletion, and in SIO all of these (except one token of *culiado*) occur exclusively with *d*-deletion.

Two of these forms, *helado* and *pesado*, are particularly interesting. Not only are they the only two with additional non-slang meanings, but they also exhibit a lower rate of *d*-deletion, which reinforces the correlation between informal meaning and informal orthography. More specifically, there are very strong correlations such that the reduced spelling is strongly preferred with the vulgar meaning, but strongly dispreferred with the non-vulgar meaning. These correlations are represented in cross-tabulation plots (cf. Gries to appear: Section 4.1.2.2) in Figure 3: observed frequencies that are larger or smaller than expected are plotted in black and grey respectively, and the physical size of the number reflects the size of the effect (based on

**Table 6** Percentage *d*-deletion among vulgar words

Form	-ado	-ao	Percentage of -ao
<i>culiado</i> ‘motherfucker’	1	106	99.07
<i>cagado</i> ‘fucked up’	14	39	73.58
<i>aweonado</i> ‘asshole’	0	45	100.00
<i>tirado</i> ‘fucked (pp.)’	16	9	36.00
<i>chingado</i> ‘fucked (pp.)’	4	1	20.00
<i>cachado</i> ‘screwed (pp.)’	0	2	100.00
Total	35	202	85.23

the residuals). Note that, for *helado* in the left panel, the Marascuilo procedure shows that ‘ice cream’ and the ambiguous meanings of *helado* do not differ from each other significantly whereas the meaning of ‘blowjob’ differs significantly from both others.

In contrast to this frequent deletion among slang and vulgar terms, most words that occurred exclusively without *d*-deletion in SIO have more formal meanings or functions: *actualizado* ‘updated’, *educado* ‘polite’, *confirmado* ‘confirmed’, *significado* ‘meaning’, *agrado* ‘(a) pleasure’, *feriado* ‘holiday’.

### 3.5.2 Vulgar words as determined in Wiktionary

Comparing the list of words tagged as vulgar in Wiktionary to the *-ado* and *-ao* forms in our corpus yielded the six matches in Table 6.

While 2077 of 10,367 non-vulgar *-ado* words occurred with deletion (20.03%), 202 of the 237 tokens of vulgar words (85.23%) occurred with *d*-deletion, which, according to a binomial test, is virtually impossible by chance ( $P < 0.001$ ).

In sum, there is not only a strong correlation of *d*-deletion with words with special pragmatic functions in general, but also a particularly strong one with vulgar words. This in turn supports our observation in the previous section about the influence of phonology on *d*-deletion. Just as the vulgar words *fuckin’* in English is hardly ever pronounced with the standard pronunciation involving word-final [ŋ] (cf. Kiesling 1998), the vulgar Spanish words here exhibit a strong dispreference for the standard pronunciation with *-ado*, testifying to the influence of phonological patterns on orthographic regularities.

## 4 Changes in Spelling: From *ch* To *x*

### 4.1 Introduction

The next feature we explore is a substitution that changes the number of graphemes representing a single phonological segment to one, resulting in a one-to-one sound-to-character ratio. This phenomenon is not uncommon in SIO: *ll* can shorten to *i* or *y* to represent [j], and *qu* can shorten to *q* or *k* to represent [k].

This spelling change of *ch* to *x* to represent the pronunciation [tʃ] is interesting in at least two respects. First, it is slightly more complicated than other such changes because *x* has a variety of phonetic values in SIO (cf. Table 2), and although it represents [ks] in standard Spanish, it has no obvious and widespread connection to [tʃ] in non-Internet Spanish. Morala (2001) speculates *ch*→*x* occurs exclusively in Spain and is potentially explicable on the basis of bilingualism with Catalan, in which *x* can represent the phonetically similar [ç]. In our corpus constructed seven years later, however, we have no trouble finding *ch*→*x* attested by users from Latin America, and we therefore expect factors besides Catalan bilingualism to be at play.

Second, we have seen above that changes of spelling are related to matters of ‘coolness’ and social group affiliation, and Mayans i Planells (2000) suggests shortenings of this kind are largely socially motivated, representing a deliberate eschewal of tradition and formality. The *ch*→*x* change would seem to be a particularly good candidate to study this because the character *x* indexes coolness in Spanish and English alike: (i) it is frequently used in supposedly hip pop culture marketing situations much like the *e* (e.g. in *e-commerce* and *e-surance*) and the *i* (e.g. in *iPod* and *iPhone*): cf., e.g. *Xbox*, *X-men*, *X-files*, *Xterra*, *xtreme*, etc. (ii) It is a character that is generally rather infrequent: it accounts for not even 0.3% of all letters of all word types in the BNC and each occurrence is therefore more noteworthy than an occurrence of, say, a *t*. (iii) It is a character that readily invokes the word *sex* because that word is among the most frequent content words with this letter: in our SIO corpus, even though it is a loan word, *sexy* is the second most

frequent word with an *x* that has not undergone *ch*→*x*.

## 4.2 Methods

In order to find instances of *ch*→*x* in our corpus, we searched for all word forms that contained *ch* and then checked each one for alternation with *x* in the corpus, although presence of alternation was not a necessary condition for inclusion. However, we considered forms that occurred only with *x* and not *ch* not to be a part of this alternation. We examined the 50 most frequent forms by combined *ch* and *x* occurrence and identified several potentially problematic types. Standalone *ch/x* as well as *ech/ex* were discarded, as the only six tokens of *ech* were found in entries written in German. Examining each individual concordance of the remaining ambiguous forms, we retained the three tokens of *chk* since they were used as a form of *chico/a* ‘boy/girl’, but revised the frequency of *xk* to zero because each token was used not as an alternation of *chk* but as a form of *por qué* ‘why’ or *porque* ‘because’ (on the basis of the pronunciation of the multiplication symbol *x* as *por*). The frequency of *xo* was revised to 2, as it occurred as an alternation of (*e*)*cho* ‘I miss’ only twice, serving in other cases as a form of *pero* ‘but’ (again on the basis of the multiplication symbol) or entry-final iconic representation of *hug and kiss*. The frequency of *xoxo* was likewise revised to 1, as it occurred only once as an alternation of *chocho* ‘cunt’. We retained all *ch* and *x* forms of *bechos* and *grachias* (i.e. *besos* ‘kisses’ and *gracias* ‘thanks’) as these can reflect an affricated pronunciation variant instead of a direct *s*→*x* or *c*→*x* orthographic alternation, and we had no reason to assume the pronunciations were not intended to be affricated. As above for *-adol/-ao*, we searched our results for English words in the manner described in Section 32 and discarded 207 types that were not also Spanish words or proper nouns. Finally, we used the same list of vulgarities as in the previous section.

While the very nature of the *-adol/-ao* deletion process determines much of the change’s phonological contexts (cf. above), the change *ch*→*x* can occur in many different places, which is why we decided to investigate to what degree the place of *ch* in the word or the syllable correlated with the rate of

**Table 7** The frequencies of *ch* and *x* in word-initial positions and elsewhere

	<i>ch</i>	<i>x</i>	Total
word-initial	13,553	7,460	21,013
elsewhere	22,244	7,961	30,205
Total	35,797	15,421	51,218

change to *x*. We decided to look at the following contrasts:

- word-initial versus elsewhere;
- pre-vocalic versus elsewhere;
- post-vocalic versus elsewhere
- intervocalic versus elsewhere;<sup>8</sup>
- hard letters (*a*, *o*, *u* before which *c* and *g* are realized as stops) versus soft letters (*e* and *i*, before which *c* and *g* are realized as fricatives).

Using the first contrast for illustration, the currently most frequent way to evaluate such data would be by means of chi-square test by generating a table such as Table 7 and then compute a chi-square test and, ideally, also an effect size measure such as  $\phi$ /Cramer’s *V*.

However, given that many movie descriptions and/or comments will contribute more than one *ch/x* spelling to these data, the chi-square test’s assumption of the independence of data points is violated (cf. Evert 2009 for detailed discussion). Evert’s (2009) recommendation is to, therefore, not make each use of *ch/x* a data point, but each description/comment. We therefore decided to compute an index for each description/comment that quantifies the degree to which it prefers *ch* or *x*, and the index we chose is the difference coefficient that has been used in, for example, Leech and Fallon (1992). Imagine a comment containing 5 word-initial forms with *ch*, and 3 with *x*. The difference coefficient is then computed as in (3):

$$(3) \frac{\text{occurrences of } x - \text{occurrences of } ch}{\text{occurrences of } x + \text{occurrences of } ch} \\ = \frac{3 - 5}{3 + 5} = -0.25$$

That is, the value of the difference coefficient ranges from  $-1$  to  $+1$ , and the smaller or larger it is, the less



**Table 8** The preferences for *ch* and *x* in word-initial positions and elsewhere<sup>14</sup>

	not word-initial			Total
	strong pref. for <i>ch</i>	no strong pref.	strong pref. for <i>x</i>	
word-initial				
strong pref. for <i>ch</i>	120,671	340	4,090	125,101
no strong pref.	243	11	54	308
strong pref. for <i>x</i>	5,726	62	1,108	6,896
Total	126,640	413	5,252	132,305

pronunciation *grachias*, it would be an instance of  $c \rightarrow x$  (although *grachias* does occur three times in the CdE).

To some degree at least, the factors governing the  $ch \rightarrow x$  alternation are the same as those for *d*-deletion. First and as before with *d*-deletion, there is a frequency effect such that the percentage of  $ch \rightarrow x$  for the most frequent words is lower than that for less frequent words, although this effect does not seem as strong here.

Second, there is also again some interaction of pragmatics and frequency at work: many of the words most frequently spelt with *x* have particular pragmatic functions:

- affectionate (though vulgar) terms of address: *wacho* ‘bastard’, *chucha* ‘cunt’, *choro* ‘vulva’, *pucha* ‘pussy’;
- forms of the farewell ‘ciao’: *chauc*, *chaus*, *chauz*, *cha*, *chauzz*, *chaoo*, *chauu*, *chao*, *chau*;
- other words with affective function: *bechos* ‘kisses’; *echo* (*de menos*) ‘I miss’ (person); *muchio*, *mcho* ‘much’;
- metadiscourse marker: *cacho* ‘I get it’.

Although discourse-pragmatic words such as the above ones make up the bulk of the top 50 forms, lower proportions of  $ch \rightarrow x$  occur in words that are not discourse-pragmatically noteworthy such as *hecho* ‘done’, *noche* ‘night’, and the proper noun *Chile*. The alternation  $ch \rightarrow x$ , as Mayans i Planells (2000) speculates, thus appears to be confirmed as a stylistic resource that is most productive in socially interactive functions. The one apparently strongest counterexample to the role of pragmatics could be the various spellings of *mucho* ‘much’, a word which is discourse-pragmatically inconspicuous. However, in this particular genre its high frequencies of

occurrence are due to its large number of occurrence in pragmatic formulae such as greetings of the kind of *cuídate mucho* ‘take care’, and *te quiero mucho* ‘I love you’, plus some greetings as in *muchos besos* ‘many kisses’.<sup>10</sup>

We also investigated whether  $ch \rightarrow x$  is more common among vulgar words. However, contrary to *d*-deletion,  $ch \rightarrow x$  turned out to not correlate with vulgarity, neither among the top 50 words represented in Figure 4 nor when we compared our data to the list of vulgarities from Wiktionary ( $\chi^2=0.8537$ ,  $df=1$ ,  $P=0.36$ ). At present, it is not clear to us why  $ch \rightarrow x$  differs from *d*-deletion this way, although the way *d*-deletion reflects the informal phonology associated with swearing in a way  $ch \rightarrow x$  does not, may provide a partial explanation.

Third, also as with *d*-deletion,  $ch \rightarrow x$  often combines with other features of SIO. *Mcho*, for example, is much more frequently affected by  $ch \rightarrow x$  than its standard counterpart *mucho* ‘much’; the same is true of *chiko* and *chika* in comparison with *chico* ‘boy’ and *chica* ‘girl’.

#### 4.4 Results 2: phonological environments

For each of the five contrasts we explored, we generated a table of the kind illustrated in Table 8 for the contrast of word-initial positions versus all other ones.

While 92% of the descriptions/comments share the same preference for  $ch/x$  (cf. the main diagonal), there is still one major difference. If one compares how the row sums, which indicate the preferences for the word-initial position, differ from the column sums, which indicate the preferences of all other positions, with a chi-square test for distribution-fitting, then the positive Pearson residuals show

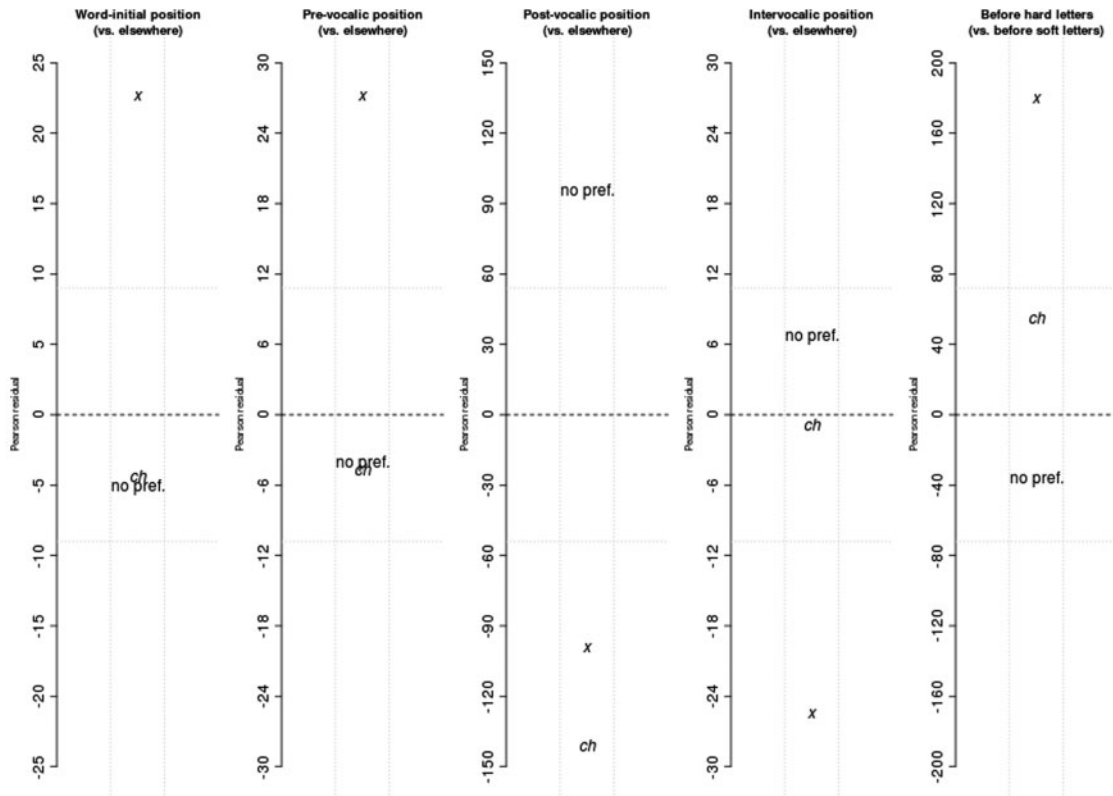


Fig. 5 Pearson residuals of *ch/x* preferences in five phonological contexts

that word-initial positions strongly prefer *x* (Pearson residual for 6,896 = 22.69) while the negative Pearson residuals for the other two rows show weaker dispreferences. Using the same kind of analysis we obtained the results represented in Figure 5.

The results are relatively clear: the rate of spelling change varies as a function of two phonological contexts such that *x* is preferred in word-initial and pre-vocalic positions, and one orthographic context such that it is preferred before hard letters.<sup>11</sup> However, after vocalic elements, no strong preference for either spelling can be observed. Although the specific implications of these trends are unclear to us, *ch* → *x* is confirmed as somewhat systematic in its distribution, illustrating how even apparently non-phonological, Internet-exclusive phenomena may still be constrained by phonology and traditional graphemics. It will be fascinating to investigate, a few years down the line, whether the same

constraints still apply or *ch* → *x* becomes more general and disperses to more environments.

## 5 Repetitions in Spelling

### 5.1 Introduction

CMC does of course not allow speakers to mark their utterances in the same way as linguistic communication does. However, since speakers do want to express states of affairs and, more importantly, attitudes and emotions, Internet language in general has evolved other ways to mark utterances. Three mechanisms are particularly frequent:

- the use of emoticons to express their general emotional stance or their emotional stance towards the propositional content of the utterance; examples include ‘:-O’ to communicate surprise or ‘:-@’ to communicate anger;

- the use of capitalization to mark (parts of) utterances as stressed and intensify their message such that *lol* represents ‘laughing out loud’ and *LOL* represents a more intensified version;
- the use of character repetition such that *lol* represents ‘laughing out loud’ and *lool* represents a more intensified version (cf. Moreno de los Rios 2001).

We will focus on the third mechanism here, character repetition, for the following reasons. First, even some exploratory eyeballing of the corpus files reveals that character repetition is extremely frequent and pervasive: we found more than 100,000 word tokens involving character repetitions, with approximately 2/3 of these being two-character sequences.

Second, a phenomenon that frequent is of course in need of explanation. Several possibilities are conceivable. One proposal is that expressions involving character repetitions—especially reduplications when the word also exists without a reduplication, *lol* and *lool* being cases in point—are actually regular abbreviations for phrases just like others without reduplication(s). For example, while *lol* stands for ‘laughing out loud’, *lool* has been suggested to stand for ‘laughing out outrageously loud’ (cf. <<http://www.netlingo.com/right.cfm?term=LOOL>>).

Another view, which we find more plausible, is to treat cases of character repetition as instances of iconically-motivated repetition (cf. Sapir 1921, p. 79). More specifically, what is probably at work here is the iconicity principle of quantity, according to which the amount of phonetic material (e.g. lengthening or reduplication of syllables) reflects quality/intensity or quantity/pluralization. Even a short stay in Internet chat rooms will show that, while reduplications account for the vast majority of character repetitions, expressions such as *loool*, *loool*, etc. are also frequently used. It is therefore much more reasonable to assume that speakers exploit a cross-linguistically attested and cognitively motivated iconicity principle than to assume that speakers use these as abbreviations for fully spelt-out expressions, especially since we have variants of *lol* involving 10 and more *o*’s.

There is one additional complication, though. If the iconicity principle of quantity were the only

factor at work here, one would expect to find all characters being repeated to the same degree (or proportionally to their overall frequency in the texts), but even a cursory glance at the frequencies with which all characters are repeated shows this not to be the case, which raises the question what it is that interacts with the iconic principle. One plausible candidate brings us back to the fact that CMC does not allow speakers to mark expressive and attitudinal stances in the same way as face-to-face conversation does. The maybe most marked difference is that CMC does not have prosody, but given that standard examples of the iconicity principle of quantity involve phonological features, character repetition may take over some aspects of what prosody does in verbal/oral linguistic communication. Obviously, for example, lengthening of syllables is exactly what, in writing/typing, would be indicated by character repetitions. This leads to two things. First, it leads to the prediction that the degrees to which characters are repeated in typing should be proportional to the degree to which their corresponding phonemes can be lengthened in speaking, maybe even more precisely to the position of the phonemes represented by the characters on the sonority hierarchy.

Second, to the degree to which this prediction is confirmed, this would constitute some *prima facie* evidence in favor of associating this variety of CMC more strongly with spoken than with written language.

We decided to distinguish three different kinds of character repetitions: namely character repetitions at the beginnings of words, words that consist of the repetition of one character only, and character repetitions at the ends of words. The reasons for this distinction are twofold. First, there is the *a priori* reason that there is psycholinguistic evidence that different parts of words are differently salient in terms of processing (cf. Noteboom 1981), and we wanted to make sure we would be able to discern different tendencies for different parts of words. Second, there is the *a posteriori* reason that we found many occurrences of whole words that consisted just of *l*’s, and very many of these were in parentheses. It turns out that this is a convention resulting from the fact that, on MSN Messenger,

**Table 9** Frequencies of word-initial repetitions per character (most frequently attested  $\approx 74.1\%$  of the data)

	length=2	length=3	length=4	length=5	...	Totals
<i>l</i>	6,126	17	2	0	5	6,150
<i>i</i>	3,731	261	21	7	14	4,034
<i>a</i>	1,213	332	67	53	178	1,843
<i>e</i>	962	213	39	19	112	1,345
<i>o</i>	734	191	100	25	229	1,279
<i>f</i>	863	5	0	0	4	872
...	3,984	360	106	49	199	4,698
Totals	17,613	1,379	335	153	741	20,221

typing ‘(L)’ produces a heart emoticon (outside of MSN, ‘(L)’ has evolved into the more iconic heart representation of ‘<3’). However, our data were not produced with MSN Messenger so speakers exploit the fact that they do not have to type exactly ‘(L)’ and intensify their message with repetitions involving two or even many more *l*’s or create new forms by omitting closing parenthesis or forming novel emoticon combinations like ‘(LL:)’. Further, *ff* was determined to often be an abbreviation for *Friends/Favorites*, a feature of one of the social networking sites. We therefore considered it prudent to distinguish different locations of repetitions in words.

## 5.2 Results 1: word-initial character repetitions

For word-initial character repetitions, we used the regular expression ‘ $\wedge(\.)\{1, \}(!\{1\} \$)$ ’ to extract all word tokens that began with a character repetition, but had additional characters after that initial repetition. This yielded 20,221 word tokens, which consisted of 4,885 word types and 273 different word-initial character repetitions.

We then first studied the distribution of the lengths of the repetitions. There is a nearly perfect inverse correlation between the logged lengths of the repetitions and the logged frequency with which they are observed ( $\tau = -0.86$ ,  $z = 7.71$ ,  $P < 0.001$ ): there are many tokens of short repetitions—the two-character sequences accounted for a large proportion of the repetitions (17,613 tokens,  $\approx 87.1\%$ )—but there were also some very long character repetitions (e.g. character repetitions involving

more than 20 characters, and in five cases even more than 50 characters).

Second, we studied the frequencies with which word-initial characters are repeated. We extracted the first characters of all word-initial repetition and then counted how often each character was part of a word-initial repetition; the characters are listed in descending order of frequency in (4). In addition, we generated a frequency table of each character and its frequency of occurrence in all repetitions. Since the resulting table is too large and too sparse to be shown here *in toto*, Table 9 provides those cells of that table that account for  $\approx 74.1\%$  of the data.

(4) *l i a e o f t j u b m s c p x r n z q d h y w k g v*

There is a very clear articulatory effect: (i) the characters that are repeated most often are continuants: one liquid, several vowels, and one fricative; (ii) the characters that are repeated most rarely are consonants and often stops; (iii) the characters that are repeated much more often than the average (e.g. are repeated 5 times or more often) are nearly exclusively vowels.

In addition to the expected articulatory effect, the fact that there are some very long character repetitions also reveals what one might call a medium effect. It is obvious that words with repetitions of 10 and more characters are not words that can be considered abbreviations of any kind. Also, these repetitions are longer than can be reasonably accounted for in terms of pronunciation lengths, and are really only genuinely possible in CMC and not in, say cell phone text messaging because repetitions of this length only arise when all the speaker

**Table 10** Frequencies of whole-word repetitions per character (most frequently attested  $\approx 64.1\%$  of the data)

	length=2	length=3	length=4	length=5	...	Totals
<i>f</i>	5,808	46	14	5	48	5,921
<i>o</i>	912	556	782	190	980	3,420
<i>i</i>	2,884	169	27	6	39	3,125
<i>l</i>	605	783	36	11	102	1,537
<i>a</i>	717	314	105	64	210	1,410
<i>n</i>	469	507	90	41	291	1,398
...	3,365	1,276	538	212	1,424	6,815
Totals	14,760	3,651	1,592	529	3,094	23,626

needs to do is hold down one key on the keyboard long enough. These aspects support the iconicity approach to character repetition discussed above. Note, however, that, so to speak, articulation constrains idiomaticity most of the time: while it is just as easy to hold a key *g* down for 5 s to produce more than 100 characters, speakers just hardly ever do that—they do this nearly only with characters representing sounds that are, in principle at least, pronounceable for a corresponding duration.

### 5.3 Results 2: whole-word repetitions

Let us now turn to words that consist of one repetition only. We followed basically the same approach and, to anticipate things, obtained very similar results. We used the regular expression  $\text{'}^{\wedge}(\.)\{1,\}\$'$  to extract all 23,626 word tokens/456 types that consist of only one repeated character.

Again, we found a Zipfian distribution such that two-character sequences accounted for large share (14,760 tokens,  $\approx 62.5\%$ ), there were few very long repetitions resulting from holding a key down long, and the lengths of the repetitions exhibited a very strong inverse correlation with their frequency ( $\tau = -0.78$ ,  $z = 8.83$ ,  $P < 0.001$ ). Also, the characters that are repeated are very much from the same set as those that account for the majority of the word-initial character repetitions; cf (5) and Table 10.

(1) *f o i l a n m e u s d k x y t b q h r p c z j g w v*

It is plain to recognize the articulatory effects again: the high repetition frequency of continuants, the low frequency of stops, and that, again, the longest repetitions are largely found with vowels only.

### 5.4 Results 3: word-final character repetitions

Finally, let us turn to word-final character repetitions. We retrieved all words that contained a final character repetition ( $\text{'}(\.)\{1,\}\$'$ ) and deleted from that all words that we previously identified as words being one repetition only. We obtained 169,570 word tokens/42,025 types with 814 word-final character repetition types. Given the findings for the other two kinds of character repetitions, these can now be discussed much more briefly as they are virtually identical: continuants are repeated most and longest (here, the effect of vowels is strongest) while obstruants are not, and we found both the overall Zipfian correlation between lengths and their frequencies ( $\tau = -0.82$ ,  $z = 12.23$ ,  $P < 0.001$ ) as well as the usual predominance of two-character sequences (113,521 tokens,  $\approx 66.9\%$ ). Both the articulatory and the medium effect are therefore further supported.

### 5.5 Results 4: a brief look at word tokens with repetitions

Let us finally have a brief look at which words are most likely to undergo changes of spelling changes. To that end, we trimmed down the repetition(s) in each word to just single instances of the repeated characters (e.g. *mucho*, *muchoo*, *muchooo* etc. were all slimmed down to *mucho*) and then determined the words that exhibited the largest number of different spelling variants. The 111 words with the largest numbers of spelling variants (13 or more) are listed in their most frequently attested spelling variant in (6).

- (1) *amoo* (44), *ee* (40), *xdd* (38), *oo* (35), *holaa* (33), *xauu* (33), *aa*, *muaa* (32), *sii* (30), *olaa*, *mm*, *xx* (29), *lll*, *ss*, *xao* (28), *byee* (27), *noo*, *tkmm*, *tqmm* (26), *uu* (25), *muchoo*, *kk*, *quieroo*, *wenaa* (24), *chao* (23), *lindaa*, *dd*, *kieroo*, *wii* (22), *saludoss*, *lindoo*, *muxoo*, *besoss*, *waa* (21), *baii*, *aiozz*, *rikoo*, *salu22* (20), *kiss*, *adoroo*, *hoolaa*, *amigaa*, *buuu* (19), *ii*, *cuidatee*, *yaa*, *adioss* (18), *ehh*, *ahh*, *chauu*, *amorr*, *xauzz*, *jaa*, *haa*, *hermosoo*, *pp*, *natyy* (17), *tuu*, *besoo*, *naa*, *pasatee*, *uii*, *vidaa*, *hoola*, *uff*, *zhaoo*, *uuii*, *woow*, *weenaa* (16), *ff*, *mii*, *nnn*, *fotoo*, *saludoos*, *pliss*, *aah*, *rikaaaaa*, *teamooo*, *weena*, *rrrrrrrrrrrrrrrr*, *grr* (15), *aii*, *diala*, *besitoss*, *mioo*, *amoor*, *pff*, *aahh*, *mass*, *woo* (14), *tee*, *see*, *shaoo*, *hermosaaa*, *tt*, *keroo*, *besitoo*, *tqqmm*, *ohh*, *mill*, *amigoo*, *cumplee*, *olii*, *ooh*, *extranoo*, *aiooss*, *olee*, *encantaa*, *aaay*, *xauss*, *zz* (13)

Given the genre of which our corpus constitutes a sample, the classes of words in which repetitions are frequent are actually not particularly surprising:

- discourse markers
- greetings: *olaa*, *hoolaa*, *hoola* ('hello'), *shaoo* ('ciao'), *aiozz*, *aiooss* ('bye'), *saludoss*, *salu22* ('greetings'), *besoss*, *besitoss* ('kisses'), plus English *baii*, *byee*, *kiss*, etc.
- emotional stance: *oo*, *woow*, *pff*, *rrrrrrrrrrrrrrrr*, *grr*, etc.
- vocalizations: *ehh*, *jaa*, *haa*, etc.
- expressions of emotions other than discourse markers: *amoo* ('I love'), *encantaa* ('enchants'), *tqmm*, *tqqmm*, *tkmm*, ('I love you'), etc.
- terms of address: *amorr*, *amoor* ('love'), *vidaa* ('life'), *amigoo* ('friend'), etc.
- *I* plus verb phrases: *quieroo*, *kieroo* ('I want'), *extranoo* ('I miss'), *adoroo* ('I adore'), etc.
- positive adjectives: *rikoo*, *rikaaaaa* ('great'), *hermosoo*, *hermosaaa* ('beautiful'), etc.

## 6 Concluding remarks

It is interesting to locate the genre studied in this article on the continuum between spoken and written data as well as with regard to the different categories of CMC as suggested by Crystal (2001).<sup>12</sup>

With regard to the former, the genre of our corpus falls somewhere between archetypal spoken and archetypal written language, which is not that interesting and would probably have been expected. It is interesting, however, in the way in which it is located in the middle.

On the one hand, in very many respects, the language in our corpus exhibits many characteristics that are archetypically written. According to Crystal's (2001, pp. 26–8) or Baron's (2003) criteria that have to do with the *circumstances of production*, our corpus of comments and descriptions involves language that is clearly space-bound, static, not immediately interactive and involving planned production, that contains punctuation and other visual structuring elements but not prosody as verbal language would etc.

On the other hand, in terms of the *function* of these texts, our corpus contains language that fulfills the typical functions of both written and spoken language since both facts or ideas about pictures and videos as well as social relations are communicated, and we have seen especially the high frequency of, and the many spelling changes in, many discourse-pragmatic markers and other pragmatically marked expressions.

Finally but most interestingly, in terms of the *linguistic means to attain the social functions*, the language of our corpus is clearly spoken. While the written medium does of course not exhibit phonological or prosodic markers of informality, emphasis, association with social groups, etc. that spoken language would employ, extralinguistic cues were often approximated (e.g. by emoticons) and each of the three phenomena exhibited, or were constrained by, decidedly phonological mechanisms, which sometimes even interacted with other linguistic factors: stress patterns (*d*-deletion), syllable structure/segmental patterns (*ch*→*x*), and manner of articulation (character/phoneme repetitions).

With regard to the latter, the classification of our Internet Spanish, the genre of our corpus defies an easy categorization given how genres in the Internet have developed. Consider Crystal's (2001, pp. 26–8) classification, which distinguishes web, email, chat, and virtual worlds. The

comments in our corpus, for example, already fall somewhere between email and chat: on the one hand, they are like email in the sense that there is often a uniquely identifiable addressee (the poster of the movie, the poster of a previous comment) and one finds intertextual reference in the form of responses to previous communication. On the other hand, they are like chat: many of them are very short, highly interactive (in the sense that they involve interjections and vocalizations), are highly context-bound and refer only to the very recent previous communication, and once they are posted, they can be read by everyone (unlike email).

A similar picture emerges when our data are compared to Baron's (2003, Figure 4) CMC spectrum: much of our data exhibits all characteristics from nearly all points on the continuum: In some sense, the descriptions and comments are available through self-archiving, they are comments and provide interaction, dialogues with and without pseudonyms are possible, and, as mentioned above, the comments as well as the interaction that they may trigger can be read by everyone.

While our study has been somewhat exploratory and descriptive (given the absence of much previous work on the topic), we hope to have shown that, not only are there very interesting patterns to be observed in Internet Spanish, but they also relate to many different levels of linguistic analysis and provide, so to speak, a snapshot on how this genre develops as we write these lines as well as how it is influenced by linguistic and cognitive processes below the level of consciousness. Of course, much more remains to be done for a language as widely used on the Internet as Spanish. First, obviously we need more descriptions as well as explanations of the above kind. Second and more importantly, one of the most pressing issues is concerned with the study of Internet language as a whole. It seems to us as if much work on register analysis, and we are again especially thinking of Biber's multidimensional approach, is not yet utilized in coming to grips with the particular properties of Internet language. Quite obviously, CMC is a genre characterized by a large degree of within-genre heterogeneity, and most classifications that have been developed so

far are not only somewhat subjective, but also leak in the sense of not doing justice to the multidimensional nature of genres in general and Internet language in particular. We therefore hope that our first more general study of Internet Spanish will be only among the first to trigger more studies pursuing the above two objectives.

## Acknowledgements

Thanks to the audience at Corpus Linguistics 2009 for feedback (to Gries and Myslín 2009). The usual disclaimers apply.

## References

- Baron, N. S.** (2000). *Alphabet to Email: How Written English Evolved and Where It's Heading*. London, New York: Routledge.
- Baron, N. S.** (2003). Why email looks like speech: proof-reading, pedagogy, and public face. In Aitchison, J. and Lewis, D. M. (eds), *New Media Language*. London, New York: Routledge, pp. 102–13.
- Baron, N. S.** (2003). Language and the internet. In Farghali, A. (ed.), *The Stanford Handbook for Language Engineers*. Stanford: CSLI Publications, pp. 59–127.
- Biber, D.** (1995). On the role of computational, statistical, and interpretive techniques in multi-dimensional analysis of register variation. *Text* 15.3:314–70.
- Cervera Rodríguez, Á.** (2001). La irrupción del coloquialismo en Internet y las nuevas tecnologías. Paper given at the Second International Conference on the Spanish Language. Valladolid, Spain.
- Crystal, D.** (2001). *Language and the internet*. Cambridge: Cambridge University Press.
- Davies, M.** (2002). *Corpus del Español (100 million words, 1200s-1900s)*. URL <<http://www.corpusdelespanol.org>>.
- Evert, S.** (2009). Rethinking corpus frequencies. Paper presented at ICAME 2009, University of Lancaster.
- Gries, St. Th.** (2009). *Quantitative Corpus Linguistics with R: A Practical Introduction*. New York: London: Routledge, Taylor & Francis Group.
- Gries, St. Th.** (in press). *Statistics for Linguistics with R: A Practical Introduction*. Berlin, New York: Mouton de Gruyter.

- Gries, St. Th. and Myslín, M.** (2009) *k dixez?* A corpus study of Spanish Internet orthography. Paper presented at Corpus Linguistics 2009, University of Liverpool.
- Internet World Stats** (2009). Internet world users by language: top 10 languages. URL <<http://www.internetworldstats.com/stats7.htm>>, accessed 12 July 2009.
- Kilgarriff, A.** (1996). BNC frequency lists. URL <<http://www.kilgarriff.co.uk/bnc-readme.html>>.
- Leech, G. and Fallon, R.** (1992). Computer corpora: what do they tell us about culture? *ICAME Journal*, **16**: 29–50.
- Llisterri, J.** (2002). Marcas fonéticas de la oralidad en la lengua de los chats: elisiones y epéntesis consonánticas. *Revista de Investigación Lingüística* 2.5:61–100.
- Mayans i Planells, J.** (2000). Género confuso: género chat. *Revista TEXTOS de la CiberSociedad*, 1. Temática Variada. <<http://www.cibersociedad.net/textos/articulo.php?art=22>>.
- Morala, J. R.** (2001). Entre arrobas, eñes y emoticones. Paper given at the Second International Conference on the Spanish Language. Valladolid, Spain.
- Moreno de los Ríos, B.** (2001). La Internet en español y el español en los mensajes electrónicos. Paper given at the Second International Conference on the Spanish Language. Valladolid, Spain.
- Piñeros, C.-E.** (2009). *Estructuras de los sonidos del español*. Upper Saddle River, New Jersey: Pearson Education.
- R Development Core Team.** (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <<http://www.r-project.org>>.
- Sapir, E.** (1921). *Language: An Introduction to the Study of Speech*. New York: Harcourt: Brace and Co.
- 2 Indeed, the high degree of cross-border social networking in the Spanish-speaking world raises questions about regional linguistic variation online: (to what extent) is internet Spanish more geographically homogeneous than other varieties of the language? Comparing the frequencies of certain regionalisms in an internet corpus versus a general corpus could give one cursory measure of this, but the question is beyond the scope of this article.
- 3 Since some data come from users in, for example, the USA, the presence of English-only entries in the corpus is not surprising. More intriguing, however, is the role of English code-switching and lexical borrowing in Spanish Internet discourse. Vulgar and pragmatically oriented words such as *bitch*, *sexy*, and *bai* ‘bye’ are frequent, suggesting English, like innovative Spanish orthography, affords some measure of covert prestige in online discourse in Spanish.
- 4 Llisterri omits *privado* from this count because it has the specific function of signaling private conversations on IRC-Hispano and thus occurs 2080 times (233 of which with *d*-deletion), while the next most frequent word occurs only 22 times. The count for our corpus reflects 13 occurrences of *privado*, one with *d*-deletion.
- 5 For this comparison, the forms of *privado* were excluded from both corpora (see previous note).
- 6 Since we expected the *-ao* variant to be infrequent in standard Spanish, all *-ao* matches in the CdE were examined and all tokens that did not qualify as instances of *d*-deletion were discarded. *Dao*, *lao*, and *nao* when used as proper nouns were discarded, as well as *nao* when used as a phonetic transcription, *nao* ‘ship’, or Portuguese *não* ‘no’. We then checked the matches of these forms in SIO, but found no tokens that did not qualify as *d*-deletion.
- 7 An interesting side remark can be made with respect to *demasiado*. This form is an apparent counterexample to the trend that words that exhibit other non-standard spelling variations are more likely to also exhibit *d*-deletion: it exhibits a non-standard *i*-reduplication but the *-ao* form is fairly infrequent. This may point to another interaction in the sense that the *i*-reduplication makes the word longer and it may thus be counterproductive (in the sense of conveying hip spellings) to then shorten the word again with the *d*-deletion. However, since we have only one example of this type, we are hesitant to formulate stronger generalizations at this point.
- 8 Pre-vocalic, post-vocalic, and inter-vocalic are defined here without overlap; thus the *ch* in *mucho*, for example, is counted only as inter-vocalic, and not also pre-vocalic and post-vocalic.

## Notes

1 We use regular expressions to characterize the patterns: sets of individual characters between square brackets are single-character alternatives; the pipe symbol ‘|’ separates (several) multi-character alternatives; the dollar sign ‘\$’ represents the end of a word; elements in regular parentheses are memorized and can be recalled with ‘\1’; the notation {*x,y*} means between *x* and *y* occurrences of the immediately preceding element.

- 9 Since we expected *ch*→*x* to be a largely web-exclusive phenomenon, all *x* matches in the CdE were examined to verify that they were in fact the result of *ch*→*x*. Several CdE tokens were discarded: all tokens of *Xica*, *Xico*, and *Xuxa* were determined to be proper nouns; *puxa* occurred exclusively in Portuguese; and *dixo* was always a phonetic transcription. As with the CdE *-ao* matches, all tokens of these words in SIO were examined but all qualified as instances of *ch*→*x*.
- 10 On the basis of a random sample, the tendency for *x* to be preferred in discourse-pragmatic uses is significant for *mucho* ( $\chi^2 = 5.1$ ;  $df = 1$ ;  $P < 0.05$ ;  $\phi/V = 0.16$ ).
- 11 After vocalic elements, no strong preference for either *ch* or *x* can be observed, which is not surprising since Spanish phonotactics does not allow for [tʃ] in coda position, and only a relatively few, non-normative forms in the SIO corpus matched the post-vocalic search expression.
- 12 We are well aware of the fact that the speaking versus writing distinction is only a rather coarse simplification, as has been shown beyond doubt in, say, Biber's multidimensional approach (cf., e.g. Biber 1995). This approach has not received the attention it deserves in overview works on internet language/CMC (cf. Baron 2003, who does not even refer to this line of work).
- 13 The use of *w*, *k*, and *sh* is particularly interesting since none of these occurs in standard spellings of non-borrowed Spanish words.
- 14 Cf. the appendix for the corresponding tables for the other phonological contexts.

## Appendix

**Table A1** Pre-vocalic versus not pre-vocalic *ch/x*

	not pre-vocalic			Total
	strong pref. for <i>ch</i>	no strong pref.	strong pref. for <i>x</i>	
pre-vocalic				
strong pref. for <i>ch</i>	2,556	8,833	307	11,696
no strong pref.	9,359	100,538	3,717	113,614
strong pref. for <i>x</i>	303	5,655	1,037	6,995
Total	12,218	115,026	5,061	132,305

**Table A2** Post-vocalic versus not post-vocalic *ch/x*

	not post-vocalic			Total
	strong pref. for <i>ch</i>	no strong pref.	strong pref. for <i>x</i>	
post-vocalic				
strong pref. for <i>ch</i>	114	212	13	339
no strong pref.	20,357	101,194	10,219	131,770
strong pref. for <i>x</i>	46	112	38	196
Total	20,517	101,518	10,270	132,305

**Table A3** Intervocalic versus not intervocalic *ch/x*

	not intervocalic			Total
	strong pref. for <i>ch</i>	no strong pref.	strong pref. for <i>x</i>	
inter-vocalic				
strong pref. for <i>ch</i>	2,566	9,025	322	11,913
no strong pref.	9,158	100,536	5,731	115,425
strong pref. for <i>x</i>	283	3,620	1,064	4,967
Total	12,007	113,181	7,117	132,305

**Table A4** *ch/x* before hard letters versus before soft letters

	before 'soft letters'			Total
	strong pref. for <i>ch</i>	no strong pref.	strong pref. for <i>x</i>	
before 'hard letters'				
strong pref. for <i>ch</i>	2,459	11,711	101	14271
no strong pref.	6,303	101,412	1,126	108,841
strong pref. for <i>x</i>	76	8,419	498	8,993
Total	8,838	121,542	1,725	132,305