“i didn't spel that wrong did i. Oops”: Analysis and normalisation of SMS spelling variation

Author's name
Author's institution

Abstract

Spelling variation, although present in all varieties of English, is particularly prevalent in SMS text messaging. Researchers argue that spelling variants in SMS are principled and meaningful, reflecting patterns of variation across historical and contemporary texts, and contributing to the performance of social identities. However, little attempt has yet been made to empirically validate SMS spelling patterns and verify the extent to which they mirror those in other texts.

This article reports on the use of the VARD2 tool to analyse and normalise the spelling variation in a corpus of over 11,000 SMS collected in the UK between 2004 and 2007. A second tool, DICER, was used to examine the variant and equivalent mappings from the normalised corpus. The database of rules and frequencies enables comparison with other text types and the automatic normalisation of spelling in larger SMS corpora.

As well as examining various spelling trends with the DICER analysis it was also possible to place the spelling variants found in the SMS corpus into functional categories; the ultimate aim being to create a taxonomy of SMS spelling. The article reports on the findings from this categorisation process, whilst also discussing the difficulty in choosing categories for some spelling variants.

Introduction

Spelling variation, although present in all varieties of English, is particularly prevalent in SMS text messaging (D. Crystal 2008) just as it is in historical varieties (such as Early Modern English), child and non-native learner language, and computer-mediated communication (such as instant messaging). Researchers argue that the choices made regarding spelling variants in SMS are functional, principled and meaningful. They are principled in the sense that

they follow orthographic principles of English and as such reflect and extend existing patterns of variation across historical and contemporary texts (T. Shortis 2007); and meaningful because they contribute to the performance of social identities (C. Tagg 2012). As yet, however, little attempt has been made to empirically validate SMS spelling patterns and verify the extent to which they mirror those in other texts.

In this paper, we report on the use of VARD2 (A. Baron & P. Rayson 2009) to manually normalise the spelling variation in CorTxt (C. Tagg 2009), a corpus of over 11,000 SMS messages collected in the UK between 2004 and 2007. A second tool, DICER (Discovery and Investigation of Character Edit Rules) (A. Baron et al. 2009a), was used to analyse the normalisations made in VARD2. DICER extracts letter replacement rules which can transform the variant form to its standard equivalent and builds a detailed database of these rules and their frequencies. Through examining DICER's analysis and categorising the spelling variants we can build a pattern of spelling trends in SMS for comparison with other text types. Our longer term aim is not only to better understand the nature of SMS spelling, but to construct a set of spelling rules which can be used to automatically normalise spelling in larger SMS corpora.

The paper is structured as follows. We start in Section 1 with an overview of research into SMS text messaging which highlights the significance of spelling variation as well as the lack of empirical validation. In Section 2 we give a brief description of VARD2 and DICER, before describing CorTxt in Section 3. The manual normalisation process and some initial findings are discussed in Section 4, with the spelling patterns found with DICER presented in Section 5. Finally, we conclude the paper and look to future research directions.

1. The significance of SMS spelling variation

The popular view of text messaging – or at least that represented in the print and online media – has tended to be that spelling variation is chaotic and likely to hinder meaning rather than create it. C. Thurlow's (2006) survey of newspaper coverage of 'issues related to young people, language and new technology' (p. 671) included 101 articles mainly from the UK and the US. Two of the recurring assertions across the articles were that 'textese' represents a 'decisive and dramatic break with conventional practice' (p. 672) and that these radically distinct practices can be linked to fallings standards in literacy. As C. Thurlow (2006, p. 677) acknowledges, it should also be pointed out that newspapers have since picked up on and reported academic work that suggests a more positive effect on literacy. Take, for example, an article entitled 'Texting boosts children's literacy ;-)' in *The Sunday Times* (May,

2008). Nonetheless, it seems fair to suggest that, to the casual reader, the media remains sceptical to some extent of such reports. To cherry pick a particularly plum example from a British newspaper, in an article from *The Guardian* in 2007 entitled 'I h8 txt msg: how texting is wrecking our language', John Humphrys writes that texters are:

> "vandals who are doing to our language what Genghis Khan did to his neighbours eight hundred years ago.
>
> They are destroying it: pillaging our punctuation; savaging our sentences; raping our vocabulary. And they must be stopped."

The sentiment is somewhat extreme but it reflects many of those captured in Thurlow's earlier study. However, as will be described shortly, these popular assumptions – that 'textese' is unique, disorganised and damaging – have been challenged by academic researchers.

As well as assuming the destructive and arbitrary nature of SMS spellings, coverage such as that mentioned above may also feed into popular beliefs as to the most common forms of spelling variation in texting, despite the lack of much evidence that such forms are actually very frequent (e.g. C. Thurlow & A. Brown 2003). In the above examples, we can point to the use of 8 in *h8* and the consonant writing (*txt msg*); as well as the emoticon (*;-)*) in the more positive report. Even D. Crystal (2008), who doubts the use of some extreme forms of abbreviation reported in the press (and suggests the benefits of texting for literacy), cannot resist perpetuating this conception of textese in the title of his popular book, *Txt Msg: the Gr8 Db8*. He also provides a list of reported English spelling variants (pp. 189-198) which include the kind of abbreviated phrases he nonetheless feels are unlikely to occur in 'real' text messages (*AFAIK, ASLMH, BION, ICWUM, PTMM, TTYL8R*).

Scholarly research based on (relatively small) datasets makes several observations which challenge the popular perspective on texting. Firstly, as mentioned above, variation in spelling is found to occur far less often than is popularly assumed. C. Thurlow & A. Brown (2003), for example, find that abbreviations occur in less than 20% of the overall message content, and other studies produce comparable figures (N. Doring, 2002; A. Deumert & S. Masinyana, 2008). Where variation in spelling does occur, however, the words most likely to be shortened are short, common ones. As R. Grinter & M. Eldridge (2003) note in their study of teenage texting practices:

> "Instead of being long challenging phrases offered by dictionaries, the teenagers recorded shortening simple words such as tomorrow and weekend, which often appeared in messages discussing plans. Other commonly shortened words included school, football, Internet, lessons, and homework."

The implications are that spelling variation emerges from use (that is, by texters abbreviating words that they commonly use) and that interlocutors are likely to understand the recurring 'respelt' versions. In other words, spelling variation is functional, principled and meaningful.

Spelling variants or 'respellings' are functional in the sense meant by T. Shortis (2007) in his discussion of 'Txt' (the language of texting) in that they emerge during interaction as a response to immediate functional demands. That is, the respellings are neither prescribed or learnt but used and picked up on in response to particular interlocutors in specific contexts. One outcome of this is that different groups will converge in practice and as such develop their own 'codes' (Tim Grant, pers comm.).[1] This sensitivity to context means that it is unlikely (although not impossible) that spelling variation would be regularly misunderstood or that it would hinder meaning between participants who are part of the same texting network or community. Integral to this is the argument that, although the tendency towards abbreviation may be shaped in part by the affordances of the technology (Y. Hård af Segerstad, 2002), texters who abbreviate are seen to respond also to communicative or 'discursive' demands (C. Thurlow and A. Brown 2003). So, abbreviations may be encouraged not only by the limited character allowance but also by the pressure to reply quickly, to use recognised forms, and perhaps to do so whilst on the move or engaged in other activities. And of course abbreviation is not the only function of spelling variation in texting. C. Thurlow and A. Brown (2003)[2] posit that 'brevity and speed' is just one motivation among three social 'maxims' which encourage variation in spelling:

a) 'brevity and speed' (seen in lexical abbreviation including letter-number homophones; and the minimal use of capitalisation, punctuation and spacing);
b) 'paralinguistic restitution' (such as the use of capitals to indicate emphasis or loudness, or multiple punctuation, which 'seeks to redress the apparent loss of such socio-emotional or prosodic features as stress and intonation')
c) 'phonological approximation' (that is, attempts to capture informal speech, which 'engenders the kind of playful, informal register appropriate to … text-messaging').

As C. Thurlow and C. Brown point out, respellings such as 'ello, goin, and bin' fulfil the need for brevity and speed as well as – in this case – approximating spoken conversation (they are examples of phonological approximation). However, in some cases, the latter two principles will override the concern with abbreviation. For example, with a form such as <nooooo!> the maxim of paralinguistic restitution requires more key presses than a simple <no> (particularly if the form is not recognised by the predictive text dictionary); as does a phonological approximation such as <nope>.[3]

Spelling variation is principled in the sense that it follows the conventions of the language's orthography, and can only be meaningful to the extent that it does so. For instance, in the case of English, to take M. Sebba's (2007) example, scratching *Down with skool* on to a school desk may be seen as a sign of rebellion (signalling a lack of education, and a disregard for the rules) – whilst scribbling *Down with sguul* would not. <Sguul> does not convey the same meaning as <skool> because it departs too far from English orthographic principles. One implication of this is that, because choices in spelling are constrained in this way, spelling variation practices in texting will reflect those in other written discourses. And, indeed, it appears that parallels can be drawn between Txt spellings and those observed in other texts. Other domains in which variation in spelling is particularly prevalent include the following:

- Historical varieties such as Early Modern English (M. Görlach 1991; A. Baron et al. 2009b).
- Advertising: such as the phonetic use of <k> in brand names such as Kwik Mart, an ongoing practice that can be traced back to the beginning of the twentieth century (L. Pound, 1925; V. Cook 2004); and the homonym <u> for you in, for example, 'High Class Shoe Repairs While U Wait' and 'Phones 4U' (V. Cook 2004).
- Dialogues in novels: eye dialect used in novels to represent a character's non-standard or regional speech and lower class or uneducated status (R. Weber, 1986; D. Crystal, 2003), such in this exchange between Tess and the 'dark queen' in T. Hardy's Tess of the d'Urbervilles (1891):

  > "How darest th' laugh at me, hussy!" she cried.
  >
  > "I couldn't really help it when toothers did," apologized Tess, still tittering.
  >
  > "Ah, th'st think th' beest everybody, dostn't, because th' beest first favourite with He just now! But stop a bit, my lady, stop a bit! I'm as good as two of such! look here here's at 'ee!"

- Pre-electronic written communication: for example, various spelling variants occurred in ordinary informal letters – A. Kesseler & A. Bergs' (2003) study of nineteenth century love letters from working class girls included the following: phonetic spellings such as <bcoz> (*because*) and <luv> (*love*); graphic symbolisations of kisses (*xx*) and roses, and grammar and spelling mistakes.
- Computer-mediated communication, such as instant messaging (R. Ling & N. Baron 2008)

It also appears that patterns of spelling variation are similar across diverse English-speaking communities. In their study of text messages written in English by iXhosa speakers in South Africa, A. Deumert & S. Masinyana (2008), for example, find respellings that reflect those identified in the US by C. Thurlow and A. Brown (2003), and posit similar motivations. Similar spelling motivations and patterns have also been documented in other languages, such as French (J. Anis, 2007).

Seen from a sociocultural perspective that posits orthography, like literacy, as a social practice (M. Sebba 2007), the fact that users are drawing on a constrained set of choices means that the respelling of words is always meaningful in the sense that it indexes particular identities. Spelling variation in SMS has been described as meaningful in the sense that it contributes to how texters portray themselves through texting, and how they position themselves in relation to the people they text. C. Tagg (2012), for example, suggests that respelling is one resource that texters also draw on in seemingly stylised ways when performing complex social identities through text messaging. Attempts at 'paralinguistic restitution' like *I'm parked next to a MINI!!!* and the 'phonological approximation' in *exercise, yeah, i kinda remember wot that is, hmm* suggest a speech-like informality which indexes a casual and intimate relationship with interlocutors (C. Tagg 2012). Also evident in SMS is a sense of play or fun. In C. Tagg's (2012) study, texters in their twenties and thirties were described as deliberately and cleverly manipulating idioms as well as spelling to convey their identities as playful and highly literate individuals: *I've just cooked a rather nice salmon a la you*. Finally, a number of other studies describe a role for spelling variation in indicating group identity and membership, particularly in reference to adolescent groups (E. Kasesniemi & P. Rautiainen, 2002; R. Ling 2005). To return to the South African study cited above, alongside what may be international Txt spellings, A. Deumert & S. Masinyana found evidence of forms which index a local identity. The message below shows how international and local forms combine.

> "Brur its 2bed one matras my darling is going 2 put me in shid in church. My money i have save have been decrease due 2 da Aunt Mayoly's funeral,& miner problst. So da case is coming very soon 3months preg. I'll c then. Sharp."
>
> (A. Deumert & S. Masinyana 2008: 126)

Locally-motivated forms (underlined above) include *Brur*, which is Afrikaans for brother; *da* for *the* (in an African-American pronunciation); and *shid*, which reflects the South African pronunciation of *shit* with a voiced stop. As such, spelling variation serves to strengthen social networks and a sense of local identity. Far from damaging language and literacy, studies suggest that 'Txt' enriches literacy practices and extends, in T. Shortis's (2007) words, the

'orthographic palette' of spelling options from which texters can draw in performing identity.

Recent research from the computational linguistics community has focussed on automatic normalisation rather than linguistic analysis of the spelling variation in SMS text. F. Yvon (2010) used finite-state transducers and R. Beaufort et al (2010) combined spell checking and machine translation approaches to normalise the orthography of French SMS messages.

In summary, the previous research presented above suggests that analysing the principles and patterns of SMS spelling variation is necessary in understanding how meaning is made and identity conveyed in text messaging. Little attempt has yet been made, however, to identify and categorise patterns of spelling variation across large datasets. It is this gap that the present study seeks to fill.

2. VARD and DICER

VARD (A. Baron & P. Rayson 2009) and DICER (A. Baron et al. 2009a) are both tools originally developed to normalise and analyse spelling variation in Early Modern English corpora (see e.g. A. Lehto et al. 2010). However, they have been adapted and trained to deal with other types of spelling variation, such as child language data (A. Baron & A. Rayson 2009) and second language learner data (P. Rayson & A. Baron 2011). For this study, VARD and DICER will be used to study SMS spelling variation found in CorTxt.

VARD[4] (version 2.3 was used for this study) functions as a pre-processor for other corpus linguistic tools by finding and inserting standard modern equivalents for spelling variants to make corpus analysis more accurate. Studies have shown that spelling variation in historical texts has a negative impact on the accuracy of key words (A. Baron et al. 2009b) and key word clusters analysis (M. Palander-Collin & M. Hakala 2011), POS tagging (P. Rayson et al. 2007) and semantic annotation (D. Archer et al. 2003). It is sensible to assume that spelling variation in other language varieties, such as SMS, is likely to have a similar impact on these corpus analysis techniques. VARD2 can be used to both manually and automatically normalise single texts or a whole corpus. The manual normalisation mode also acts as a training mechanism, with the tool learning how to best use its normalisation techniques to deal with the spelling variation in a given corpus. For each normalisation made, an XML tag is inserted into the saved text with the original form saved as an attribute, for example:

<normalised orig="l8r">later</normalised>

This format allows most corpus linguistic tools to use the normalised form in their analysis, whilst the original spelling is maintained. This pairing of the

spelling variant to the chosen normalised form can also be used to provide data to the second tool used in the analysis produced for this study – DICER. The specific character differences between the variant and normalised forms are analysed by DICER and spelling rules are created; e.g. for the above example, the rule "Substitute '8' for 'ate'" would be created. How often these spelling rules are needed for the spelling normalisations in a corpus, and in which positions[5] and context within the variants, are collated into a database and presented in a series of webpages[6] for analysis. The rules created and their frequencies are useful for the investigation of spelling trends (as presented below), but also for finding "letter replacement rules" which can be added to VARD2 to improve its normalisation performance (see A. Baron et al., 2009a).

3. Data and participants

CorTxt, the corpus used in this study, was collected for the purposes of linguistic analysis by C. Tagg (2009). The corpus specifications are shown in Table 1. Friends and family of the researcher submitted messages they had received as well as those they had sent and so the corpus includes texts from an extended 'network' which reaches beyond the researcher's own interlocutors to texters in various places across Britain. At the same time, findings from this extended network must be generalised with care (as with any texting group): the corpus is rapidly dating, contributors were older than in many studies, they comprised generally well-educated professionals or students, and with one or two exceptions they had English as their home language. In particular, it is possible that the spelling practices of this relatively literate group may involve a higher degree of sophistication than, for example, in text messages composed by adolescents. As described above, we expect that many of the spelling variants will reflect general patterns; at the same time, some spellings will emerge from the practices of these particular groups and this will have implications for the training of a normalisation tool.

No. of messages	11,067
No. of words	190,516
Average no. of words per message	17.2 words
Collection period	March 2004 – May 2007 (3 years, 2 months)
Collection methods and procedures	From Friends and Family (10,626 messages) AOL anonymous online public forum (441). A further 441 messages came from an anonymous public forum provided by AOL for forwarding text messages. After it was discontinued, people kept using it: that is, messages used in this study were those which were sent to the

	service but not forwarded.
Composition of texters	Mainly British English speakers, aged 19 – 68, professionals and students Male = 41%; F = 59%
Language	English (mainly British English)
Type of communication	Mainly personal communication, although some business text messages evident, including: *Hi Are you interested in working taster day at Cornwall College, Camborne 27.4.06 4hrs@£6 per hour? Ring if interested. NAME173*

Table 1 - CorTxt description.

4. Manual normalisation

In order to produce patterns for DICER to analyse, VARD 2.3 was used to manually normalise 2,430 randomly selected messages (41,342 tokens) from CorTxt, equating to around a fifth of the whole corpus. The randomly selected messages were split between the three authors for normalisation, except for 5% of the total messages which were included in all three sample sets in order to calculate inter-rater agreement scores. From this overlap, 200 normalisations were made by at least one person. 75% of these normalisations were fully agreed, in that all three agreed (a.) that a normalisation should be made and (b.) on the word to normalise to. Of the 50 normalisations not fully agreed:

- In 28 instances, only one person decided to make a normalisation.
- In 19 instances, two decided to make a normalisation (and agreed on the word to normalise to), the other person made no normalisation.
- In the 3 other instances, a normalisation has been made by all three people, but two have chosen a different word to normalise to than the other person.

Hence, when at least two people have made a normalisation (172 cases), 98.26% have agreement on which word to normalise to, and when all three have agreed to make a normalisation (153 cases), 98.04% have agreement on the word to normalise to. Therefore, we can conclude that overall there was very high inter-rater agreement on actual normalisations made, with a small amount of disagreement on whether or not some words should be normalised.

In total, the normalisation procedure resulted in 3,166 spelling variant normalisations, meaning just 7.66% of the tokens in the sample were considered spelling variants and normalised with VARD. Furthermore, it was found that around half of the messages (1,217) contained no spelling variants

at all. These low quantities of spelling variation confirm previous smaller-scale studies which found that spelling variation occurs far less often than is popularly assumed (N. Doring 2002; C. Thurlow & A. Brown, 2003; A. Deumert & S. Masinyana 2008).

Of the 3,166 spelling variants normalised, 322 (10.17%) could be considered "real-word errors", i.e. the variant spelling matches an unintended dictionary word. This figure could be considered quite low compared to previous studies highlighting real-word error rates; for example, J. Pedler & R. Mitton (2010) found that 31.4% of dyslexic writers' spelling errors were real-words and P. Rayson & A. Baron (2011) found that 21.9% of second language learner errors were real-words. However, as real-word errors are inherently difficult to detect automatically, this low figure requires further verification.

5. DICER Analysis

DICER was used to analyse the 3,166 spelling variant normalisations produced. The frequencies of the spelling normalisations examined resemble a common Zipfian distribution for word frequencies, with some highly frequent spelling variants found and a long tail of unique occurrences. There are 744 spelling normalisation types, with 525 being unique in our sample.[7] The 10 most frequent spelling normalisations are given in Table 2.

Variant	Normalisation	Frequency
u	you	666 (21.04%)
2	to	187 (5.91%)
r	are	95 (3.00%)
4	for	85 (2.68%)
ur	your	75 (2.37%)
tomo	tomorrow	74 (2.34%)
its	it's	71 (2.24%)
b	be	69 (2.18%)
c	see	52 (1.64%)
im	i'm	43 (1.36%)

Table 2 - Most common spelling variant normalisations.

The DICER analysis provides various results which show the spelling trends found in our sample and highlight how difficult automatic normalisation is for this type of spelling variation. For instance, only 28.5% (42% for types) of spelling variants could be normalised by a single character edit.[8] This is much lower than similar rates found for common spelling errors, e.g. R. Mitton (1987) noted that 69% of spelling errors from a range of sources could be corrected with just a single character edit, while earlier studies noted higher

rates still. Indeed, many spelling correction methods rely on this by finding dictionary words which are one edit away from a given spelling error (e.g. K. Church & W. Gale 1991). The reason why so many of our spelling variants require more than one character edit may be due to a high number of different forms of shortening. This view is supported by the high number of normalisations which require insertion of characters; 72% (57% for types) of normalisation operations[9] were insertions (27% were substitutions and less than 1% were deletions), compared to 33% of normalisations requiring insertion in a study by R. Mitton (2008: Table 3).[10] This will be discussed further in our analysis of spelling variation categories.

In terms of specific normalisation rules found by DICER, the most frequently required rules (for variant tokens) largely reflect the most common spelling variants found, as shown in Table 2. The ten most frequent rules are shown in Table 3, with all but *Insert g* and *Insert space* relating directly to the top 10 most common spelling variants (see, for example, insert 'yo'), although most rules also incorporate other variants, particularly *Insert '* and *Insert e*. In other words, the normalisation rules appear to be shaped by the very frequent occurrence of particular variants, chiefly <u>, rather than the general application of a particular rule across variants. Considering only each individual spelling variant once, the top 10 most frequent rules in terms of types are given in Table 4. There is quite a lot of overlap between the two top 10 lists, although *Insert space* is the most frequent for types (clearly due to the possibility to omit spaces between potentially any combination of words). *Insert o* and *Insert h* are also present in the types top 10, but not for tokens. As previously mentioned, these rules not only highlight the spelling trends found in our sample, but provide possible rules which could be used to assist in spelling normalisations both manually and automatically. The rules which appear most frequently for tokens will help with normalising the most frequent spelling variants found (e.g. those given in Table 2). On the other hand, rules that appear most frequently for types will help with normalising a larger variety of spelling variants.

Rule	**Frequency**	**Top position**	**Examples**
Insert yo	802 (21.67%)	Start (100%)	u, ur, u'd
Insert '	387 (10.46%)	Penultimate (72.61%)	its, dont, thats
Insert e	351 (9.48%)	End (87.46%)	r, b, hav
Substitute 2 → to	249 (6.73%)	Start (99.60%)	2, 2day, 2moro
Insert a	162 (4.38%)	Start (91.98%)	r, n, bout
Insert g	122 (3.30%)	End (98.36%)	goin, comin, mornin
Insert space	109 (2.95%)	Middle (81.65%)	Thankyou, aswell, togo
Substitute 4 → for	89 (2.40%)	Start (100%)	4, 4ward, 4got

Insert rrow	74 (2.00%)	End (100%)	Tomo
Substitute c → see	54 (1.46%)	Start (100%)	c, cing

Table 3 - Top ten most frequent normalisation rules in terms of tokens.

Rule	**Frequency**	**Top position**	**Examples**
Insert space	97 (10.43%)	Middle (83.51%)	ifnot, outthatway, togo
Insert e	62 (6.67%)	End (64.52%)	pls, wher, lv
Insert ’	61 (6.56%)	Penultimate (70.49%)	cant, didnt, hes
Insert g	56 (6.02%)	End (96.43%)	mornin, lookin, usin
Substitute 2 → to [11]	34 (3.66%)	Middle (97.06%)	want2go, need2say, what2do
Insert o	17 (1.83%)	Second/Middle (35.29%)	cud, dnt, hw, anymre
Insert yo	16 (1.72%)	Start (100%)	u, uve, u’ll
Substitute 2 → to	16 (1.72%)	Start (93.75%)	2day, 2mora, 2gether
Insert h	16 (1.72%)	Second (43.75%)	wen, wich, wat
Insert a	15 (1.61%)	Start (53.33%)	bout, ne (any), n (and)

Table 4 - Top ten most frequent normalisation rules in terms of types.

1.1. Categories of spelling variation

The DICER analysis was also used to assist in categorising variants, the ultimate aim of which would be to create a taxonomy of SMS spelling variants. The categories were identified by drawing on a range of existing taxonomies which were themselves compiled with reference to various written and electronic texts (e.g. J. Androutsopoulos 2000; D. Crystal 2003; R. Weber 1986; T. Shortis 2007), and adapted to account for the spellings seen in CorTxt. As will be described below, categorisation remains a somewhat subjective process, in which the views of the researcher are to some extent imposed on the data.

The functional categories of spelling variants which we identified using DICER are given in Table 5 along with examples and the frequency of each category in terms of tokens and types (percentages are out of the number of category assignments). Note that a substantial number of forms are placed in more than one category: <2nite>, for example, includes both a number homophone <2> and the eye dialect spelling of *night* <nite>.

Category	**Examples**	**Tokens**	**Types**
Letter	u, r, ur, c, b	1040	30 (3.25%)

homophones		(29.91%)	
Number homophones	person2die, 2gether, up4that,in2hospital, 2nite	476 (13.69%)	126 (13.64%)
Clippings	tomo, tho, v, bout, prob, hav	414 (11.91%)	113 (12.23%)
Apostrophe omission	wots, im, il, its, thats	367 (10.56%)	60 (6.49%)
Eye dialect	bak, luv, wots, gud	243 (6.99%)	47 (5.09%)
Colloquial contractions	lookin, av, cos, n, whaddya	232 (6.67%)	94 (10.17%)
Spacing	Thankyou, ur, u2, aswell,Ohdear, sleep4aweek	232 (6.67%)	171 (18.51%)
Consonant writing	txt, msg, lv, wld, pls	130 (3.74%)	51 (5.52%)
Mistyping	your (for you're), definately, adn, menas	61 (1.75%)	47 (5.09%)
Double letter reduction	stil, wory, spel, I'l, 2moro, ul	43 (1.24%)	16 (1.73%)
Misspelling	your (for you're), definately,	32 (0.92%)	26 (2.81%)
Unclear	ur = your, tomoz = tomorrow	31 (0.89%)	21 (2.27%)
Other abbreviations	no, happng, checkd, 2morw	29 (0.83%)	17 (1.84%)
Possible regional respellings	summat, summort, sumfing, dis	28 (0.81%)	8 (0.87%)
Predictive texting 'mistake'	in (for go), he (for if)	6 (0.17%)	2 (0.22%)
Visual morphemes	I'm@my; Lunch@12	5 (0.14%)	5 (0.14%)
No category assigned		108 (3.11%)	90 (9.74%)
Total		**3477**	**924**

Table 5 - Category examples and frequencies.

At first glance, the functional categories may seem primarily motivated by the maxim of brevity and speed. For example, representing *you* with either <u> and <ya> involves abbreviation in the sense of fewer letters. However, the very fact that texters can choose between <u> or <ya> suggests that motivations other than brevity and speed may also be involved. In this case, the choice of <ya> (rather than <you> or <u>) represents an informal, spoken version of *you*, while the potential effects of <u> rest on the visual appearance of the word. The fact that spelling variants can be categorised into functional categories suggests meaningful and deliberate usage, a point supported by the relative infrequency of forms categorised as misspellings or mistypings.

By far the most frequently occurring category, in terms of tokens, is that of letter homophones, which account for nearly a third of all spelling variants categorised by DICER, and which include <u> (*you*), <r> (*are*), <b> (*be*) and <c> (*see*). It is important to note, however, that the first homophone, <u>, accounts for the overwhelming majority of letter homophones. This may reflect the frequency of the second person pronoun; *you* is the most frequent word in the corpus, with 7884 occurrences, 3043 of which are spelt <u> (C. Tagg 2009: 135). In other words, the frequency of this category of spelling variation may reflect the occurrence of one particular lexical item as much as it does a general tendency towards using letter homophones – the lower frequency in terms of types provides more evidence for this view. Similarly, number homophones are the second most frequent category (albeit with half the occurrences of letter homophones), and these chiefly comprise the numbers <2> and <4>. Again, this may be because of the frequency of these lexical items (*to* and *for*) which occur respectively in 2nd and 9th position in the word frequency list (and of course <2> can also represent *too*) (C. Tagg 2009: 219). What is also interesting about the number homophones is the number of times that they occur in alphanumeric sequences such as *up4that*, that is, where spaces are omitted around the homophone. The omission of spacing is a frequent occurrence throughout the corpus. Omitting spacing around number homophones rather than, say, letter homophones, is probably a strategy motivated by intelligibility. The high frequency of these alphanumeric sequences also explains why the frequency of this category in terms of types is at a similar percentage to that of tokens, unlike for letter homophones. The high frequency of <2> and <4> certainly increases the category token frequency, but the variety of words which surround <2> and <4> accounts for the high type frequency. In the light of the media portrayal, it is interesting to note that there are only 4 occurrences of <8> as a homophone in the sample, all in the sequence *l8r*. As mentioned at the start of this paper, <8> appeared from our selected extracts to be a feature popularly associated with SMS (as in Humphrys' 'I h8 txt msgs').

The next most frequent category is clippings, accounting for nearly 12% of the respellings. A clipping is defined by D. Crystal (2003) as 'part of the word which serves for the whole, such as ad and phone': that is, where either the beginning or end of the word has been clipped (or the middle in cases such as <spectacles> clipped to <specs>). One could speculate that these, at least where the end of the word is clipped, may have become more common with the advent of predictive texting. The dictionary would not automatically recognise <tmwr>, but by four presses it would have predicted that the texter intended to write *tomorrow* and would suggest <tomo>. Consonant writing is in fact much less frequent in the corpus. Whether speed would be achieved alongside brevity depends on whether predictive texting is used, and whether the vowel-less form had been entered in the dictionary. If predictive texting

was used, and the form was not in the dictionary, it is likely that the variant form would involve more key presses than the 'standard' form. The assumption here, of course, is that brevity and speed are the motivations.

With colloquial contractions and eye dialect, both of which account for just under 7% of all spelling variant tokens, texters appear to be combining the maxim of brevity and speed with that of phonological representation; that is, they are capturing informal spoken forms. To take the (slightly) less frequent of the two first, colloquial contractions were defined by R. Weber (1986: 415) as 'spelling patterns that are regularly used to represent reduced, colloquial speech'. These include <kinda>, <ya>, <lookin> and <cos>. By definition, these forms involve abbreviation. However, the spellings also capture informal spoken forms. The other category, eye dialect, is in fact more complex than a simple representation of forms. Respelt forms such as <gud> and <wots> do not reflect how words would be pronounced in fast, informal speech but represent a more common sound-symbol relationship (for example, <o> for the vowel sound in *what*). The effect of these is often to suggest an uneducated, ignorant speaker (D. Preston 2000). In SMS, then, although the forms are generally shorter, the motivation for their use may lie in conveying a certain kind of identity (rebellious, careless, unconcerned) rather than simply an attempt to reduce characters. If colloquial contractions and eye dialect can be described as indexical, the implication is that homophones, clippings and consonant writing may similarly index identities, albeit in a way less directly associated with speech.

Before moving any further down the list of categories, it is useful to pause and consider how cautiously should the above assertions be made? As mentioned earlier, identifying functional categories and assigning certain forms to them relies to an extent on our assuming what texters were doing or what they thought they were doing. In some cases, forms could have been placed in more than one category, and the category that we chose to assign it to is therefore arguable. These include the examples shown in Table 6 (in each case, the chosen category comes first).

Spelling	**Category Choice**	**Normalisation**
<b>	letter homophone – or clipping?	(be)
<n>	colloquial contraction – or clipping?	(and)
<tho>	clipping – or eye dialect?	(though)
<prob>	clipping – or colloquial contraction?	(probably and problem)
<bout>	clipping – or colloquial contraction?	(about)
<bak>	eye dialect – or abbreviation?	(back)
<lookin>	colloquial contraction – or clipping?	(looking)
<wiv>	colloquial contraction – or regional spelling?	(with)
<checkd>	other abbreviations – or consonant writing?	(checked)
<happng>	other abbreviations – or consonant writing?	(happening)

<tomoro>	clipping – or eye dialect?	(tomorrow)

Table 6 - Examples were category choice was particularly debatable.

Some of these raise wider issues. To what extent can regional representations be distinguished from 'standard' colloquial contractions? The spellings <summort> and <sumfing> seem fairly marked forms particular to certain dialects (e.g. West Country and London, respectively); as does <dis> for *this*. The rendering of *with* as <wiv> however, is more contentious, as is the reduction of <ing> to <in> and the haitch-dropping in <av> (for *have*). There is likely no absolute dividing line between these two categories. It could also be argued that 'double letter reductions' such as <stil> and <spel> are phonetic representations and could be categorised as eye dialect.

Some spellings were more difficult to put into any category at all. These included <tomoz> (for *tomorrow*) and <ur> (for *your*). The latter is particularly interesting. The spelling <ur> is used throughout the corpus to mean either *you are* or *your*. The former can be understood as a pair of letter homophones (<u> for *you* and <r> for *are*). This explanation does not hold for *your* represented as <ur>. A more accurate explanation may be that <ur> is a clipping of *your*. However, this explanation feels less satisfactory than a phonetic one, perhaps given the phonetic use of <ur> elsewhere. If <ur> represents *you are*, and therefore *you're*; then it seems a small step to representing *your* as <ur>. Given the uncertainty, this variant was placed in the Unclear category (as was <tomoz>).

The results reveal very few forms which were deemed to be 'mistakes', either mistypings (such as <adn> for *and*) or misspellings (such as <definately>). This has some implications for the category of apostrophe omission, which occurs very frequently across the corpus. Given that spelling variation elsewhere in the corpus is seemingly so deliberate and principled, one could argue that apostrophe omission may similarly constitute deliberate attempts to either save time, space or effort, or to create a certain effect. In the case of <wots> for example, the lack of apostrophe adds to the eye dialect to create a sense of rushed, casual speech. The main point to emerge from the relative lack of misspellings and mistypings is that spelling practices can be described as deliberate and meaningful.

There are of course a number of problems with the above assertion. One question that emerges is the extent to which 'mistakes' and deliberate respellings can be distinguished. Our starting assumption was that spelling variation is functional, principled and meaningful and this inevitably shaped our categorisation. The spelt forms <wot>, <wory> and <hav> were therefore labelled as the deliberate spelling strategies of, respectively, eye dialect, double letter reduction and clipping. Whether or not any instances of these forms were genuine mistakes is hard to tell. Coming at this from the opposite direction, those that were classified as 'mistakes' were deemed so because

they did not fit the categories we had identified – they included transpositions such as <adn> (*and*) and <menas> (*means*), as well as <your> (*you're*) and <definately> (*definitely*). However, given a different categorisation system, who is to say that transpositions could not take on some distinct meaning? The classification process is somewhat circular. This is also well illustrated by the omission of apostrophes, which occurred on numerous occasions throughout the corpus. To an extent, <wots>, <im> and <thats> can be seen as part of the intimate, casual and playful language of Txt, and one could argue that it is unlikely that the people who contributed to the corpus would have mistakenly omitted the apostrophe in, say, *that's*; but where is the line drawn? Confusion between *it's* and *its* is arguably fairly widespread and such confusion is likely to occur in this corpus.

1.2. Summary of the DICER categories analysis

Bearing in mind the caveats highlighted, certain conclusions can be drawn from the DICER analysis. Our study confirms and extends R. Grinter & M. Eldridge's (2003) observation that spelling variants tend to be of common words – they suggest *tomorrow* and *homework* among their teenage texters. Our findings reveal that very frequent grammatical words are in fact those that have their spelling most commonly altered. The most frequently varied words are *you* and *to*, and this explains the predominance of letter and number homophones <u> and <2>. In other words, a few, very commonly used and regularly respelt words account for a great deal of the variation in spelling in the corpus. On the other hand, missing are the long abbreviated phrases listed by D. Crystal (2008): *AFAIK, ASLMH* and so on. What this suggests is a different set of practices than that described in the popular press; practices which emerge from and facilitate interaction.

The identification of functional categories facilitated by DICER show that patterns of spelling variation can be categorised across large datasets. The implication is that spelling variation is motivated not only by the need for brevity and speed. As C. Thurlow & A. Brown (2003) note, spelling variation is also motivated by the communicative demands of the medium as well as the attempt to index informal and casual social identities through the representation of spoken forms (such as colloquial contractions and eye dialect). However, the indexical potential of respelling does not appear to lie only in its relation to spoken forms. Letter and number homophones, clippings and consonant writing play instead with the visual form of the word. The distribution of frequencies across the categories in our study suggests the visual may play a much larger role than attempts to reflect speech.

Spelling variants that do not fit functional categories are surprisingly few (given concerns in the print press regarding the chaotic and damaging nature

of texting), thus offering some support to the general suggestion that spelling variants are deliberate and meaningful. Overall, what emerges is that SMS spelling variants *can* be categorised across a large dataset not only according to formal patterns but also to functional categories including colloquial contractions and eye dialect.

Conclusions

In this paper, we have reported on the use of VARD2 and DICER to manually normalise and analyse spelling variation in a corpus of SMS text messages, CorTxt. Our aim was to contribute to a description of the spelling variation that occurs in text messaging with findings based on a significant quantity of empirical data. We have highlighted various spelling trends found in the corpus which expose the difficulty of automatic normalisation and uncovered specific spelling rules which allowed us to categorise the spelling variants. Our analysis suggests that a few frequently occurring forms, chiefly the homophone <u>, account for a great deal of the spelling variation, and that these are used alongside colloquial contractions and eye dialect in playful identity construction. While our research provides valuable empirical evidence for the ideas about texting being advanced by linguists, it should be noted that the findings apply to the spelling practices of an educated, literate network and that younger or less educated texters may diverge more widely from the orthographic principles described above. Further research is needed to explore the extent to which this is the case.

As detailed in the article, we were able to build up a set of spelling rules which could subsequently be used to automatically normalise spelling across larger SMS corpora. The caveat here, of course, is that spelling variation practices may differ slightly across different communities and thus datasets. In an initial study (not detailed here) we find that these rules are useful when used within VARD and allow for a substantial amount of automatic normalisation of CorTxt after training. Future work could build upon this initial study to assess how useful a spelling rule based method could be for normalising SMS corpora.

The method of spelling variation analysis described here, i.e. using VARD and DICER, could also be applied to other language varieties which contain substantial spelling variation – both for finding spelling trends and categories and for finding spelling rules which can be used for automatic normalisation. For example, the progress made in automatically normalising Early Modern English spelling variation (see A. Lehto et al. 2010) can be improved further with an improved rule base (A. Baron et al. 2009a) and other forms of computer-mediated communication (CMC, e.g. Twitter) corpora could have

their spelling variation analysed, with trends and categories found and comparisons made between CMC varieties.

Acknowledgments

Alistair Baron's and Paul Rayson's contribution to this paper was co-funded by the EPSRC and ESRC (EP/F035438/1, EP/F035071/1, EP/F035454/1), Project title "Isis: Protecting children in online social networks".

References

Androutsopoulos, Jannis K. 2000. Non-standard spellings in media texts: The case of German fanzines. *Journal of Sociolinguistics*, 10(4), 419-438.

Anis, Jacques. 2007. Neography: Unconventional spelling in French SMS. In *The Multilingual Internet: Language, culture, and communication online*, B. Danet and S.C. Herring (eds). Oxford: Oxford University Press, pp. 87-115.

Archer, Dawn; Tony McEnery; Paul Rayson; Andrew Hardie. 2003. Developing an automated semantic analysis system for Early Modern English. In *Proceedings of Corpus Linguistics 2003*, D. Archer, P. Rayson, A. Wilson, T. McEnery (eds.). Lancaster, UK: Lancaster University.

Baron, Alistair; Paul Rayson. 2009. Automatic standardisation of texts containing spelling variation: how much training data do you need? In *Proceedings of Corpus Linguistics 2009*, M. Mahlberg, V. González-Díaz, C. Smith (eds.). Liverpool: University of Liverpool.

Baron, Alistair; Paul Rayson; Dawn Archer. 2009a. Automatic standardization of spelling for historical text mining. In *Proceedings of Digital Humanities 2009*. Maryland, USA: University of Maryland.

Baron, Alistair; Paul Rayson; Dawn Archer. 2009b. Word frequency and key word statistics in historical corpus linguistics. *International Journal of English Studies* 29(1): 41-68.

Beaufort, Richard; Sophie Roekhaut; Louise-Amélie Cougnon; Cédrick Fairon. 2010. A hybrid rule/model-based finite-state framework for normalizing SMS messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.* Sweden: Uppsala.

Church, Kenneth W.; William A. Gale. 1991. Probability scoring for spelling correction. *Statistics and Computing* 1(2): 93–103.

Cook, Vivian. 2004. *Accommodating Broccoli in the Cemetery: or why can't anybody spell?* New York: Touchstone.

Crystal, David. 2003. *The Cambridge Encyclopedia of the English Language*. Cambridge: Cambridge University Press.

Crystal, David. 2008. *Txtng: the Gr8 Db8*. Oxford: Oxford University Press.

Deumert, Ana; Sibabalwe O. Masinyana. 2008. Mobile language choices — The use of English and isiXhosa in text messages (SMS): Evidence from a bilingual South African sample. *English* 29(2): 117-147.

Döring, Nicola. 2002. "1 bread, sausage, 5 bags of apples I.L.Y" - communicative functions of text messages (SMS). *Zeitschrift für Medienpsychologie* 3.

Görlach, Mandfred. 1991. *Introduction to Early Modern English*. Cambridge: Cambridge University Press.

Grinter, Rebecca E.; Margery Eldridge. 2003. Wan2tlk?: Everyday Text Messaging. In *Proceedings of the CHI'03 Conference on Human Factors in Computing Systems*. New York: ACM.

Hård Af Segerstad, Yvla. 2002. *Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication*. Ph.D. thesis, Department of Linguistics, Göteborg University, Sweden.

Hardy, Thomas. (1891) *Tess of the d'Urbervilles*. London: Random House.

Kasesniemi, Eija-Liisa; Pirjo Rautiainen. 2002. Mobile culture of children and teenagers in Finland. In *Perpetual Contact: mobile communication, private talk, public performance,* J. Katz, M. Aakhus (eds.). Cambridge: Cambridge University Press.

Kesseler, Angela; & Alexander Bergs. 2003. Literacy and the new media: vita brevis, lingua brevis. In *New Media Language*, J. Aitchison, D. Lewis (eds.). London: Routledge.

Lehto, Anu; Alistair Baron; Maura Ratia; Paul Rayson. 2010. Improving the precision of corpus methods: The standardized version of Early Modern English Medical Texts. In *Early Modern English Medical Texts: Corpus description and studies*, I. Taavitsainen, P. Pahta (eds.). Amsterdam: John Benjamins.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting insertions, deletions and reversals. *Cybernetics and Control Theory* 10(8): 707–710.

Ling, Rich; Naomi S. Baron. 2008. Text messaging and IM: linguistic comparison of American college data. *Journal of Language and Social Psychology* 26(3): 291-298.

Ling, Rich. 2005. Mobile communications vis-à-vis teen emancipation, peer group integration and deviance. In *The Inside Text: social perspectives on SMS in the mobile age*, R. Harper, A. Taylor, L. Palen (eds.), Vol. 4, *The Kluwer International Series on Computer Supported Cooperative Work*. Amsterdam: Springer.

Mitton, Roger. 1987. Spelling checkers, spelling correctors and the misspellings of poor spellers. *Information Processing and Management* 23(5): 495–505.

Mitton, Roger. 2008. Ordering the suggestions of a spellchecker without using context. *Natural Language Engineering* 15(2): 173–192.

Palander-Collin, Minna; Mikko Hakala. 2011. Standardizing the Corpus of Early English Correspondence (CEEC). Poster presented at *ICAME 32*, Oslo, 1-5 June 2011.

Pedler, Jennifer; Roger Mitton. 2010. A large list of confusion sets for spellchecking assessed against a corpus of real-word errors. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (eds.). European Language Resources Association (ELRA), Valletta, Malta.

Pound, Louise. 1925. The Kraze for "K". *American Speech* 1(1): 43-44.

Preston, Dennis R. 2000. Mowr and mowr bayud spellin': Confessions of a sociolinguist. *Journal of Sociolinguistics* 4(4): 614-621.

Rayson, Paul; Alistair Baron. 2011. Automatic error tagging of spelling mistakes in learner corpora. In *A Taste for Corpora*: *In honour of Sylviane Granger*, F. Meunier, S. De Cock, G. Gilquin, M. Paquot (eds.). Amsterdam: John Benjamins.

Sebba, Mark. 2007. *Spelling and Society: the culture and politics of orthography around the world*. Cambridge: Cambridge University Press.

Shortis, Tim. 2007. Revoicing Txt: Spelling, Vernacular Orthography and 'Unregimented Writing'. In *The Texture of Internet: Netlinguistics in Progress*, S. Posteguillo, M. J. Esteve, M. L. Gea-Valor (eds.). Newcastle: Cambridge Scholars Publishing.

Tagg, Caroline. 2009. A corpus linguistic study of SMS text messaging. Unpublished PhD thesis, English Department, University of Birmingham (March 2009).

Tagg, Caroline. 2012. *The Discourse of Text Messaging: analysis of SMS communication*. London: Continuum.

Thurlow, Crispin; Alex Brown. 2003. Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online*, 1(1).

Thurlow, Crispin. 2006. From statistical panic to moral panic: The metadiscursive construction and popular exaggeration of new media language in the print media. *Journal of Computer-Mediated Communication* 11: 667-701.

Weber, Rose-Marie. 1986. Variation in spelling and the special case of colloquial contractions. *Visible Language*, 20(4): 413-426.

Yvon, Francois. 2010. Rewriting the orthography of SMS messages. *Natural Language Engineering* 16(2): 133-159.

---

[1] Dr Tim Grant's project at the Centre for Forensic Linguistics at Aston University involved participants contributing text messages online. He used a 'snowballing technique' whereby participants were recruited through existing participants, thus enabling differences between 'networks' to emerge (see http://www1.aston.ac.uk/lss/subject-areas/english/activities/texting-study/).

[2] Thurlow and Brown's article appeared in *Discourse Analysis Online* and as an online journal there are no page numbers.

[3] In this paper, we follow the convention of marking orthographic representations with <> (see, e.g., Sebba 2007), so that *you* is a word which in texting is often spelt <u> or <ya>.

[4] For further details about how VARD2 functions, see Baron & Rayson (2009) and the software's website (http://ucrel.lancs.ac.uk/vard), where it can also be downloaded free for academic use.

[5] A position of *First*, *Second*, *Middle*, *Penultimate* or *End* is recorded.

[6] See http://corpora.lancs.ac.uk/dicer

[7] A spelling normalisation is considered unique when the combination of spelling variant and normalisation is unique.

[8] An edit here is an insertion, deletion or substation of one character, as used in Levenshtein Distance (Levenshtein, 1966).

[9] There is often more than one normalisation operation per spelling variant required, e.g. <dnt> requires the operations *Insert o* and *Insert '* to arrive at the normalised form *don't.* In total there were 3,701 normalisation operations and 930 when each variant-normalisation pair was considered once (i.e. for types).

[10] The table given by Mitton shows "Omissions" and "Insertions", these are in terms of how the misspelling is changed from the correct word, i.e. an omission means a letter is missing from the misspelling. The results from DICER are in terms of how the misspelling (or variant) needs to change to successfully correct (or normalise) it.

[11] Here the rule is substitute *2* with *to* surrounded by white space.