"I didn't spel that wrong did i. Oops" Analysis and standardisation of SMS spelling variation

Caroline Tagg, Alistair Baron and Paul Rayson

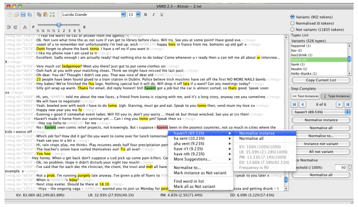
ICAME 2010, May 27th, Giessen

Outline of study

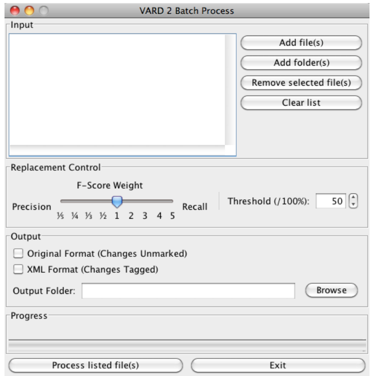
SMS Language

I didn't spel that wrong did i. Oops

CorTxt



Manually Standardised



VARD 2 Batch Process

Input

Replacement Control

F-Score Weight: Precision ⅕ ¼ ⅓ ½ 1 2 3 4 5 Recall

Threshold (/100%): 50

Output

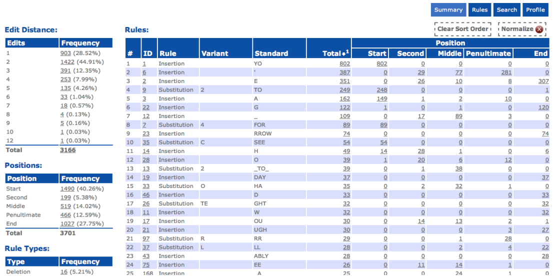
□ Original Format (Changes Unmarked)

□ XML Format (Changes Tagged)

Output Folder:

Progress

Automatically Standardised



Edit Distance:

Edits	Frequency
1	903 (28.52%)
2	1422 (44.91%)
3	391 (12.35%)
4	253 (7.99%)
5	135 (4.26%)
6	33 (1.04%)
7	18 (0.57%)
8	4 (0.13%)
9	5 (0.16%)
10	1 (0.03%)
12	1 (0.03%)
Total	3166

Positions:

Position	Frequency
Start	1490 (40.26%)
Second	199 (5.38%)
Middle	519 (14.02%)
Penultimate	466 (12.59%)
End	1027 (27.75%)
Total	3701

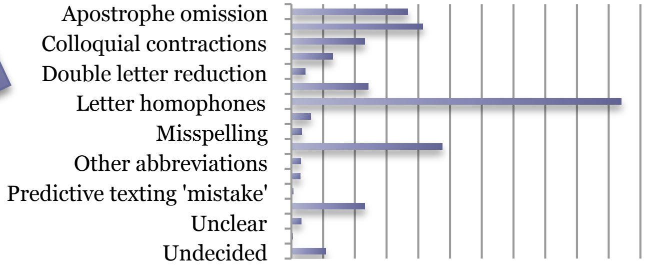
Rule Types:

Type	Frequency
Deletion	16 (5.21%)

Rules:

						Position				
#	ID	Rule	Variant	Standard	Total↓1	Start	Second	Middle	Penultimate	End
1	1	Insertion		YO	802	802	0	0	0	0
2	6	Insertion		'	387	0	29	77	281	0
3	2	Insertion		E	351	0	26	10	8	307
4	9	Substitution	2	TO	249	248	0	0	0	1
5	3	Insertion		A	162	149	1	2	10	0
6	22	Insertion		G	122	1	0	1	0	120
7	12	Insertion		_	109	0	17	89	3	0
8	7	Substitution	4	FOR	89	89	0	0	0	0
9	23	Insertion		RROW	74	0	0	0	0	74
10	35	Substitution	C	SEE	54	54	0	0	0	0
11	14	Insertion		H	49	14	28	1	0	6
12	28	Insertion		O	39	1	20	6	12	0
13	13	Substitution	2	_TO_	39	0	1	38	0	0
14	19	Insertion		DAY	37	0	0	0	0	37
15	33	Substitution	O	HA	35	0	2	32	1	0
16	46	Insertion		D	33	0	0	0	0	33
17	26	Substitution	TE	GHT	32	0	0	0	0	32
18	11	Insertion		W	32	0	0	0	0	32
19	17	Insertion		OU	30	0	14	13	2	1
20	21	Insertion		UGH	30	0	0	0	3	27
21	97	Substitution	R	RR	29	0	0	1	28	0
22	37	Substitution	L	LL	28	0	0	2	4	22
23	43	Insertion		ABLY	28	0	0	0	0	28
24	75	Insertion		EE	26	0	11	14	1	0
25	168	Insertion		A	25	0	0	24	1	0

DICER Analysis



SMS Taxonomy

Popular view of spelling variation
(Thurlow 2006)

"I h8 txt msgs: How texting is wrecking our language"

The Daily Mail, 2007

AFAIK, ASLMH, BION, ICWUM, PTMM, TTYL8R

from Crystal 2008 *Txtng: the gr8 db8*

The view in the SMS literature

- Not as common as you'd think (Doring 2002; Thurlow and Brown 2003).
- Functional, principled and meaningful (Shortis 2006) (skool vs sguul)
- Beneficial for literacy (Plester et al)
- Reflective of patterns elsewhere

> 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..

(Deumert and Masinyana 2008)

CorTxt

	Text message corpus (*CorTxt*)
No of messages	11,067
No of words	190,516
Collection period	March 2004 – May 2007
Collection method	From friends and family
No. and composition of texters	235 British English speakers, aged 19-68, professionals and students F = 62%; M = 28%

see Tagg (2009)

Alan says we can come to your birthday meal. Where will it be? Laura can stay at mine if your squashed at yours

Ok that would b lovely, if u r sure. Think about wot u want to do, drinkin, dancin, eatin, cinema, in, out, about... Up to u! Wot about NAME408? X

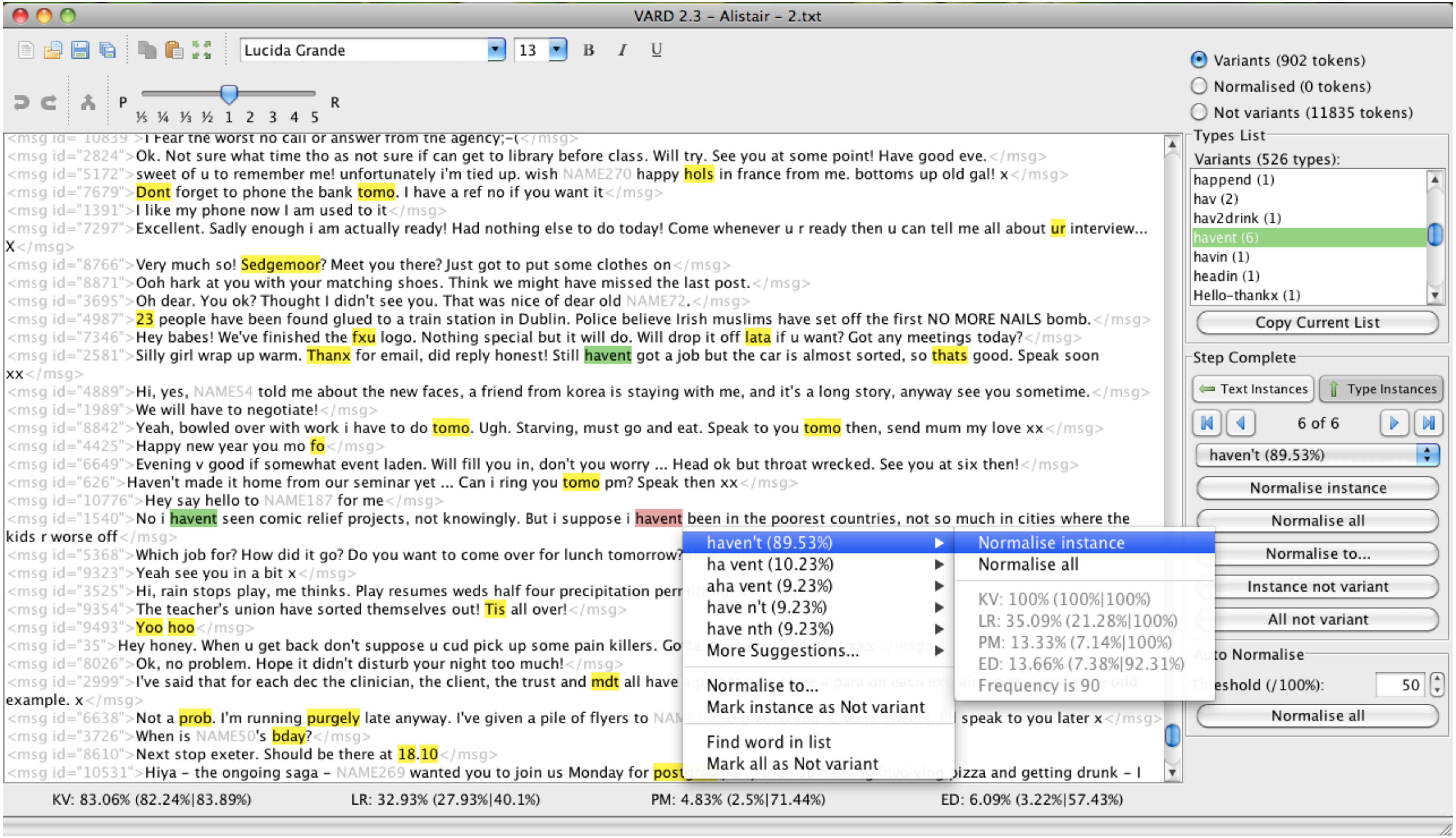
Kinda. First one gets in at twelve! Aah. Speak tomo xx

Thankyou for ditchin me i had been invited out but said no coz u were cumin and u said we would do something on the sat now i have nothing to do all weekend i am a billy no mates i really hate being single

(CorTxt)

VARD 2.3

- Originally developed to deal with spelling variation in Early Modern English.
- Can be trained to deal with any type of spelling variation.
- Functions as a pre-processor for other corpus linguistic tools to make analysis more accurate.
  - e.g. Key Word Analysis (Baron et al, 2009), POS tagging (Rayson et al, 2007) and Semantic analysis (Archer et al, 2003).
- Retains original spelling for future analysis.
  - <normalised orig=“l8r”>later</normalised>
- Freely available for academic use:
  - http://www.comp.lancs.ac.uk/~barona/vard2/



● Variants (902 tokens)
○ Normalised (0 tokens)
○ Not variants (11835 tokens)

Variants (526 types):

happend (1)
hav (2)
hav2drink (1)
havent (6)
havin (1)
headin (1)
Hello-thankx (1)

<msg id="10839">I fear the worst no call or answer from the agency;-(</msg>
<msg id="2824">Ok. Not sure what time tho as not sure if can get to library before class. Will try. See you at some point! Have good eve.</msg>
<msg id="5172">sweet of u to remember me! unfortunately i'm tied up. wish NAME270 happy hols in france from me. bottoms up old gal! x</msg>
<msg id="7679">Dont forget to phone the bank tomo. I have a ref no if you want it</msg>
<msg id="1391">I like my phone now I am used to it</msg>
<msg id="7297">Excellent. Sadly enough i am actually ready! Had nothing else to do today! Come whenever u r ready then u can tell me all about ur interview... X</msg>
<msg id="8766">Very much so! Sedgemoor? Meet you there? Just got to put some clothes on</msg>
<msg id="8871">Ooh hark at you with your matching shoes. Think we might have missed the last post.</msg>
<msg id="3695">Oh dear. You ok? Thought I didn't see you. That was nice of dear old NAME72.</msg>
<msg id="4987">23 people have been found glued to a train station in Dublin. Police believe Irish muslims have set off the first NO MORE NAILS bomb.</msg>
<msg id="7346">Hey babes! We've finished the fxu logo. Nothing special but it will do. Will drop it off lata if u want? Got any meetings today?</msg>
<msg id="2581">Silly girl wrap up warm. Thanx for email, did reply honest! Still havent got a job but the car is almost sorted, so thats good. Speak soon xx</msg>
<msg id="4889">Hi, yes, NAME54 told me about the new faces, a friend from korea is staying with me, and it's a long story, anyway see you sometime.</msg>
<msg id="1989">We will have to negotiate!</msg>
<msg id="8842">Yeah, bowled over with work i have to do tomo. Ugh. Starving, must go and eat. Speak to you tomo then, send mum my love xx</msg>
<msg id="4425">Happy new year you mo fo</msg>
<msg id="6649">Evening v good if somewhat event laden. Will fill you in, don't you worry ... Head ok but throat wrecked. See you at six then!</msg>
<msg id="626">Haven't made it home from our seminar yet ... Can i ring you tomo pm? Speak then xx</msg>
<msg id="10776">Hey say hello to NAME187 for me</msg>
<msg id="1540">No i havent seen comic relief projects, not knowingly. But i suppose i havent been in the poorest countries, not so much in cities where the kids r worse off</msg>
<msg id="5368">Which job for? How did it go? Do you want to come over for lunch tomorrow?
<msg id="9323">Yeah see you in a bit x</msg>
<msg id="3525">Hi, rain stops play, me thinks. Play resumes weds half four precipitation perm
<msg id="9354">The teacher's union have sorted themselves out! Tis all over!</msg>
<msg id="9493">Yoo hoo</msg>
<msg id="35">Hey honey. When u get back don't suppose u cud pick up some pain killers. Go
<msg id="8026">Ok, no problem. Hope it didn't disturb your night too much!</msg>
<msg id="2999">I've said that for each dec the clinician, the client, the trust and mdt all have
example. x</msg>
<msg id="6638">Not a prob. I'm running purgely late anyway. I've given a pile of flyers to NA speak to you later x</msg>
<msg id="3726">When is NAME50's bday?</msg>
<msg id="8610">Next stop exeter. Should be there at 18.10</msg>
<msg id="10531">Hiya - the ongoing saga - NAME269 wanted you to join us Monday for post pizza and getting drunk - I

KV: 83.06% (82.24%|83.89%) LR: 32.93% (27.93%|40.1%) PM: 4.83% (2.5%|71.44%) ED: 6.09% (3.22%|57.43%)

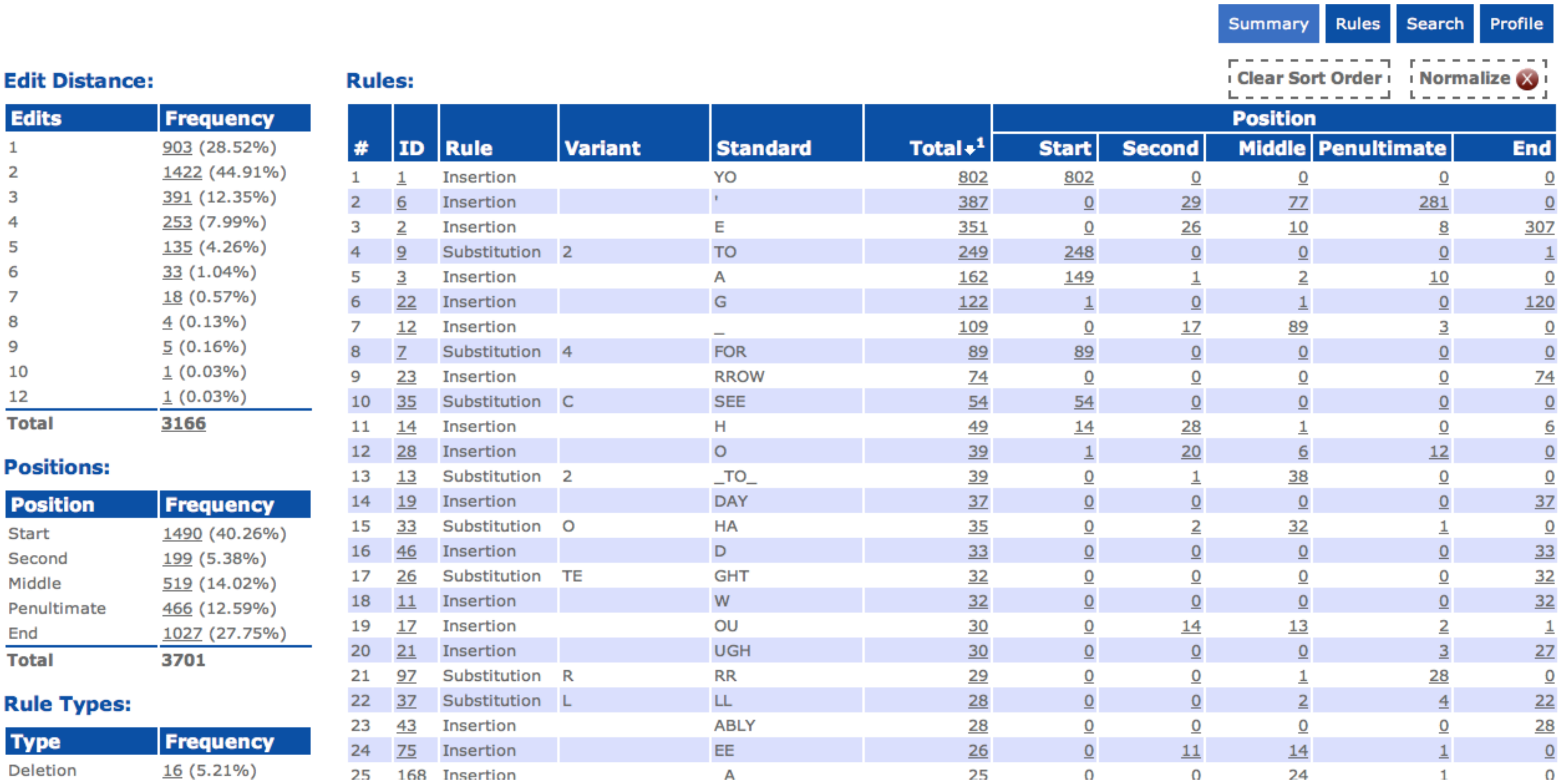
Manual Standardisation

- Around a fifth of CorTxt messages were picked at random.
  - 2430 messages.
  - 41342 words.
  - Average message length: 17 words.
  - Range from “0” to 192 words.
- Standardised with VARD 2’s interactive mode.
  - 3166 words standardised.
    - 1.3 variants per message.
    - 1217 messages contained no spelling variants.
  - 322 standardised words were “real word errors”.
  - 963 additional words marked as variants incorrectly.

DICER

- Analyses VARD output to produce letter replacement rules:
  - <normalised orig="l8r">later</normalised>
  - Rule: 8 -> ate (Middle)
- Frequencies for each rule and its context are stored in a database and are viewable in a series of webpages:
  - http://corpora.lancs.ac.uk/dicer/
- Can be plugged back into VARD 2 to improve standardisation performance.

DICER



Edit Distance:

Edits	Frequency
1	903 (28.52%)
2	1422 (44.91%)
3	391 (12.35%)
4	253 (7.99%)
5	135 (4.26%)
6	33 (1.04%)
7	18 (0.57%)
8	4 (0.13%)
9	5 (0.16%)
10	1 (0.03%)
12	1 (0.03%)
Total	**3166**

Positions:

Position	Frequency
Start	1490 (40.26%)
Second	199 (5.38%)
Middle	519 (14.02%)
Penultimate	466 (12.59%)
End	1027 (27.75%)
Total	**3701**

Rule Types:

Type	Frequency
Deletion	16 (5.21%)

Rules:

#	ID	Rule	Variant	Standard	Total↓1	Position				
						Start	Second	Middle	Penultimate	End
1	1	Insertion		YO	802	802	0	0	0	0
2	6	Insertion		'	387	0	29	77	281	0
3	2	Insertion		E	351	0	26	10	8	307
4	9	Substitution	2	TO	249	248	0	0	0	1
5	3	Insertion		A	162	149	1	2	10	0
6	22	Insertion		G	122	1	0	1	0	120
7	12	Insertion		_	109	0	17	89	3	0
8	7	Substitution	4	FOR	89	89	0	0	0	0
9	23	Insertion		RROW	74	0	0	0	0	74
10	35	Substitution	C	SEE	54	54	0	0	0	0
11	14	Insertion		H	49	14	28	1	0	6
12	28	Insertion		O	39	1	20	6	12	0
13	13	Substitution	2	_TO_	39	0	1	38	0	0
14	19	Insertion		DAY	37	0	0	0	0	37
15	33	Substitution	O	HA	35	0	2	32	1	0
16	46	Insertion		D	33	0	0	0	0	33
17	26	Substitution	TE	GHT	32	0	0	0	0	32
18	11	Insertion		W	32	0	0	0	0	32
19	17	Insertion		OU	30	0	14	13	2	1
20	21	Insertion		UGH	30	0	0	0	3	27
21	97	Substitution	R	RR	29	0	0	1	28	0
22	37	Substitution	L	LL	28	0	0	2	4	22
23	43	Insertion		ABLY	28	0	0	0	0	28
24	75	Insertion		EE	26	0	11	14	1	0
25	168	Insertion		A	25	0	0	24	1	0

DICER – Some findings

- 40% of edits required occurred at the start of the words. This is much higher than other types of spelling variation.
- 37% of rules are "Insertion". Again, much higher than other forms of spelling variation.
- 70.5% of spellings require more than one edit (insertion, deletion or substitution) to reach an equivalent standard form.

Top 10 Rules

1. Insert "yo" (start)
2. Insert apostrophe (penultimate)
3. Insert "e" (end)
4. Sub "2" -> "to" (start)
5. Insert "a" (start)
6. Insert "g" (end)
7. Insert space (middle)
8. Sub "4" -> "for" (start)
9. Insert "rrow" (start)
10. Sub "c" -> "see" (start)

DICER Categories

- New functionality added to website to allow the categorisation of spelling variants.
- Aim is to create a taxonomy of SMS orthography.
- Similar efforts have been manually produced for other computer based media:
  - Blogs and forum data (Tavosanis, 2007)
  - Instant messaging (Varnhagen et al, 2009)
- The DICER analysis can be used to assist in categorising spelling variants.

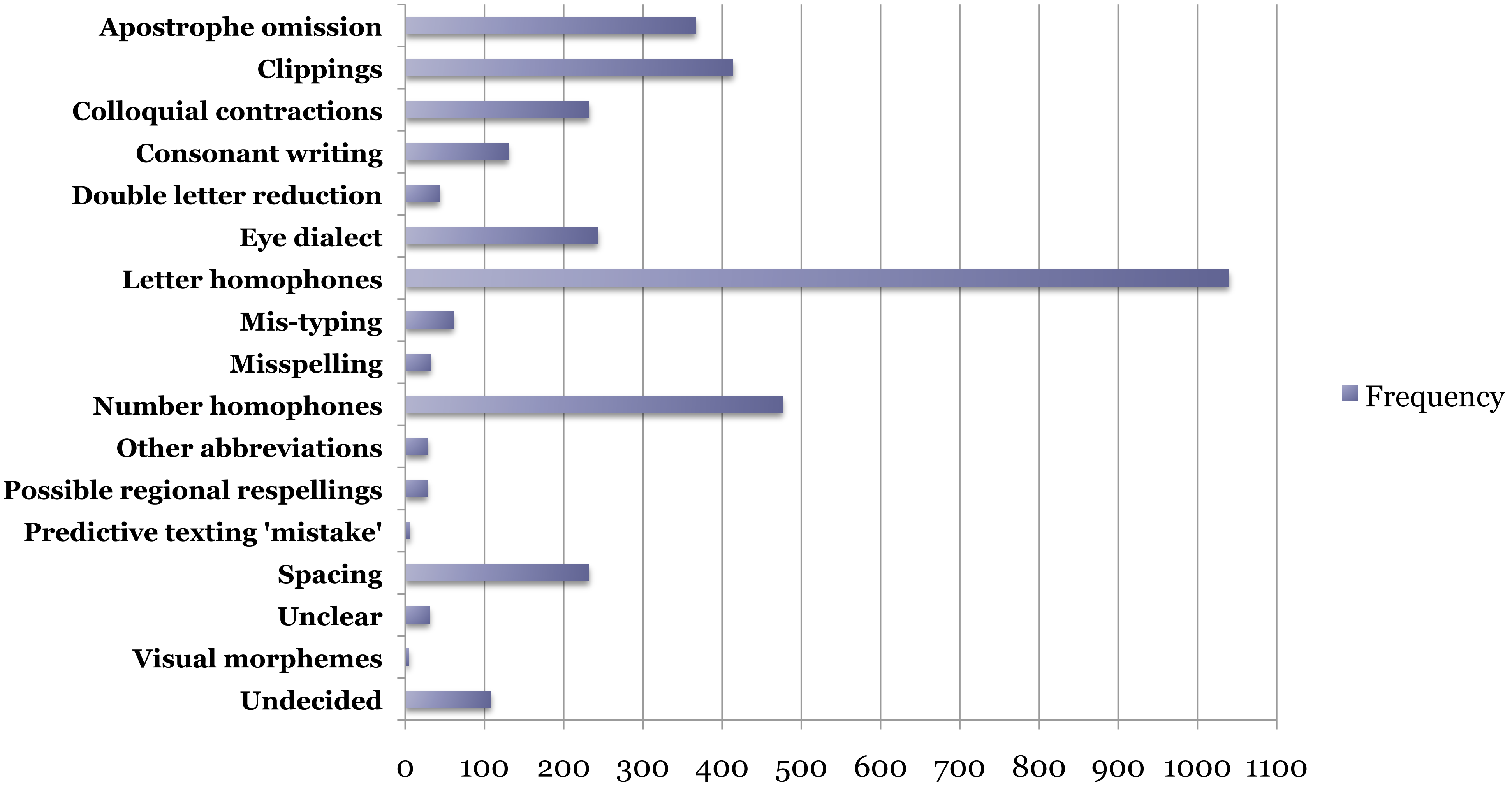
DICER Categories

• Clippings:	*tomo, tho, v, bout, prob, hav*
• Letter homophones:	*u, r, ur, c, b*
• Number homophones:	*person2die, 2gether, up4that, in2hospital, 2nite*
• Eye dialect:	*bak, luv, wots, gud*
• Colloquial contractions:	*lookin, av, cos, n, whaddya*
• Mis-spellings / -typings:	*your, definately, adn, menas*
• Unclear:	*ur = your;* *tomoz/tomoro = tomorrow*

DICER Categories

• Apostrophe omission:	*wots, im, il, its, thats*
• Consonant writing:	*txt, msg, lv, , wld, pls*
• Double letter reduction:	*stil, wory, spel, I’l, 2moro, ul*
• Other abbreviations:	*no, happng, checkd, 2morw*
• Regional respellings:	*summat, summort, sumfing, dis*
• Predictive texting mistake:	*in (for go), he (for if)*
• Spacing	*Thankyou, ur, u2, aswell,* *Ohdear, sleep4aweek*
• Visual morphemes	*I’m@my; Lunch@12*

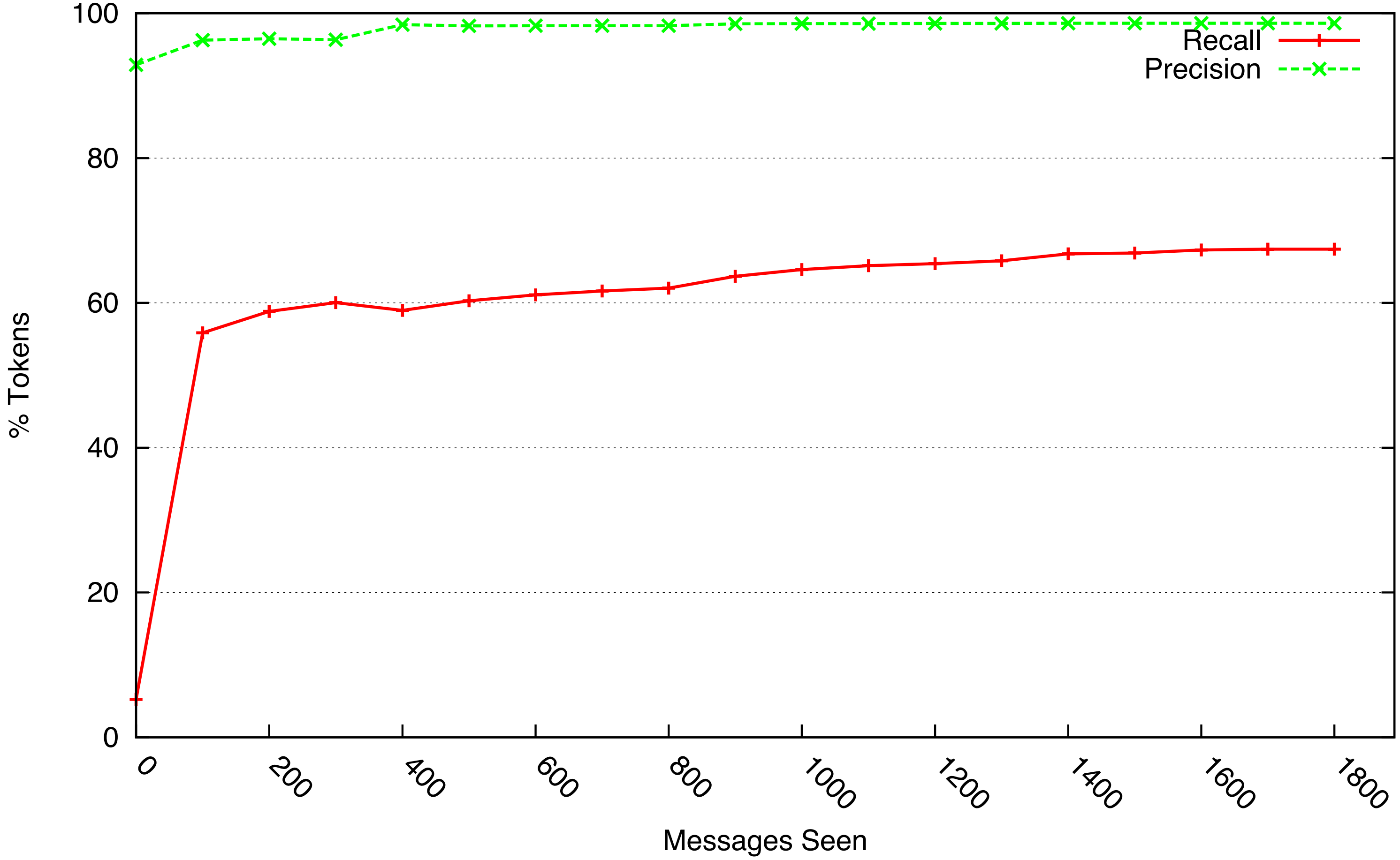
DICER Category Assignments



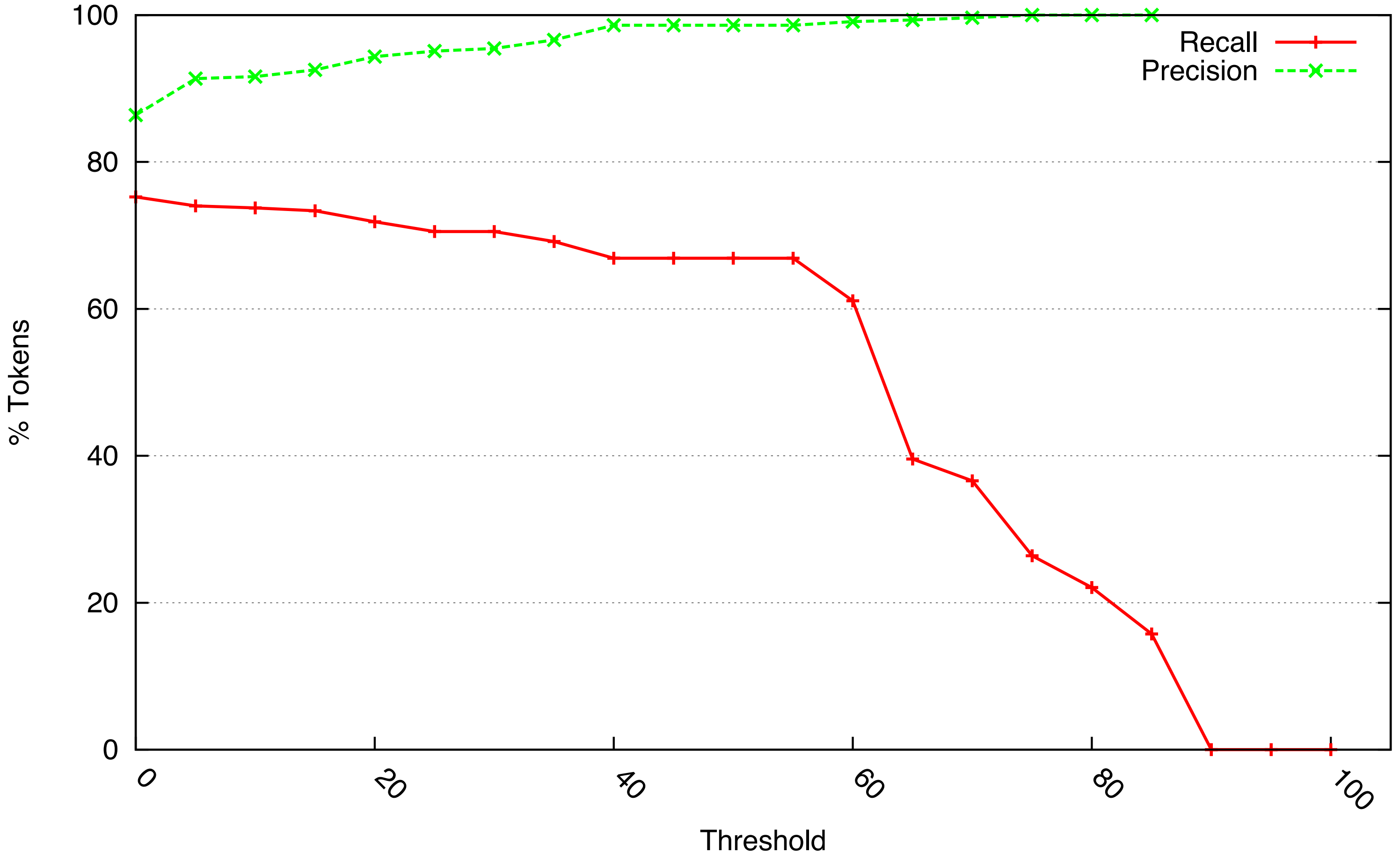
Automatic Standardisation

- Manually standardised samples split into 4 equal parts. 3 parts for training, 1 part for testing.
- Letter replacement rules were added from the DICER analysis.
  - ▫ Minimum frequency of 10.
  - ▫ Contexts of each rule was taken into account.
- The known variants list was discarded before training.

Automatic Standardisation: Training



Automatic Standardisation: Replacement Threshold



Conclusions

- SMS spelling variation is principled and meaningful.
- DICER facilitates the categorisation of these spelling decisions.
- SMS spelling throws up different challenges for standardisation.
- Nonetheless, VARD 2 can still accurately standardise a large portion of SMS spellings.

Acknowledgements

- Thanks to Paul for his help in the study and manual standardisation.

- Alistair and Paul's contribution is part of the Isis Project:
  - "Protecting children in online social networks"
  - 3 year EPSRC/ESRC funded project.
  - Lancaster, Swansea, Middlesex and specialist UK law enforcement agencies.
  - http://www.comp.lancs.ac.uk/isis/

References

Archer, D., T. McEnery, P. Rayson and A. Hardie (2003). "Developing an automated semantic analysis system for Early Modern English." *Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper number 16*: 22-31.

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

Baron, A., Rayson, P. and Archer, D. (2009). Word frequency and key word statistics in historical corpus linguistics. In Ahrens, R. and Antor, H. (eds.) *Anglistik: International Journal of English Studies*, 20 (1), pp. 41-67.

Deumert, A. and S. O. Masinyana (2008) 'Mobile language choices - the use of English and isiXhosa in text messages (SMS): evidence from a bilingual South African sample' *English World-Wide* 29/2: 117-147.

Rayson, P., D. Archer, A. Baron, J. Culpeper and N. Smith. "Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora." *Proceedings of Corpus Linguistics 2007*. University of Birmingham, UK (27-30 July 2007).

References

Plester, B., C. Wood and P. Joshi (2009) 'Exploring the relationship between children's knowledge of text message abbreviations and school literacy outcomes' *British Journal of Developmental Psychology 27/145-161.*

Shortis, T. (2007) 'Gr8 Txtpectations: the creativity of text spelling' *English Drama Media Journal* 8/21-26.

Tagg, C. (2009) A Corpus Linguistics Study of Text Messaging. PhD thesis, University of Birmingham.

Tavosanis, M. (2007). "A causal classification of orthography errors in web texts". In *Proceedings of AND 2007.*

Thurlow, C. and A. Brown (2003) 'Generation Txt? Exposing the sociolinguistics of young people's text-messaging' *Discourse Analysis Online* 1/1.

Thurlow, C. (2006) 'From Statistical to Moral Panic: the metadiscursive construction and popular exaggeration of new media language in the print media' *Journal of Computer-Mediated Communication* 11/3: 667-701.

Varnhagen, C., G. Mcfall, N. Pugh, L. Routledge, H. Sumida-Macdonald, and T.Kwong (2009). "lol: new language and spelling in instant messaging". *Reading and Writing*, Online First.