Recognizing and Converting Cockney Rhyming Slang
for Cyberbullying and Crime Detection

Radoslaw Komuda[†], Michal Ptaszynski[††], Rafal Rzepka[†] and Kenji Araki[†]

[†] Hokkaido University, Sapporo
{komuda,rzepka,araki}@ist.hokudai.ac.jp

[††] Kitami Institute of Technology, Kitami
ptaszynski@cs.kitami-it.ac.jp

Abstract

This paper presents a Cockney rhyming slang recognizing and converting modules of a cyberbullying detection system. Firstly, we introduce the concept of rhyming slang, analyze its phrasal constructions and discuss the usefulness of features of the rhyming slang, such as resemblance to code-mixing. Secondly, we describe the corpus and phrasal rhyming lexicon created for the purpose of the research and present the results of the experiments on recognizing and transforming rhyming slang constructions into casual English sentences. Finally, we process the obtained output and verify usability of the modules in cyberbullying and crime detection system.

1 Introduction

Electronic communication has changed the way we communicate. In this sense it also changed the way we live in general. The increased tempo of distributing, sharing and receiving new information has almost a directly proportional reflection in our daily matters – people do not want to wait [Chen *et al.*, 2005]. This impatience in combination with our dependence on technology and its rapid development has switched us from thinking "if" something is possible to "when" will it be available.

Possibilites of new technologies bring new challenges to the world. However, not all of them are as new as one would assume. A good example of this problem would be cyberbullying (CB), or bullying and slandering people on the Internet, which has become a major social problem negatively influencing mental health of, mostly young, Internet users. The main goal of this research is to support CB detection efforts by creating a tool for universal slang conversion and testing it on Cockney Rhyming Slang (CRS).

In section 2 we discuss cyberbullying as a (rediscovered) social problem and recall research result presenting its scale. Section 3 introduces selected examples of the solutions addressing the problem of cyberbullying and the importance of slang issues, especially for the British society. In section 4 we present CRS lexicon and construcion creation details. Section 5 covers the matters of lexicon and corpus creation and entries categorization. Experiment results on CRS recognition and convertion as well as cyberbullying (CB) and crime detection in CRS with error analysis can be found in Section 6 and Appendix 1. Final section presents conclusions and future work plans.

2 Challenge of Cyberbullying Detection

As a social problem, cyberbullying has emerged with the rapid development of network and mobile technologies which has given a new life to well-known problems: threats, blackmail, sexual harrasement, mocking, abuse, distribution of false and embarrasing information to humiliate othersor stalking.

The novelty within these problems has to be examined in two perspectives. First of all, each and all of the challenges has expanded its areas of influence (horizontal perspective). To compare a classic bullying with its cyber equivalent we can note the expansion of the literal area itself – persecution is no longer limited to a school or work enviroment. With space comes time independency – new technologies give aggressors a 24/7 access to their victims [Cyberbullying Awareness] and are more permanent. Secondly, new technologies equip the online harassers with new ways to assault their victims (vertical perspective): directly (e-mails, instant messages) and publicly (Internet forums, Facebook pages); with offensive words, embarrassing images or photomontages.

It has been reported in numerous studies that the problem of cyberbullying can affect as much as 20% of young people

[Hasebrick *et al*., 2009; Li, 2007; Pyzalski, 2012; Kann *et al*., 2014] though usuall estimates across Europe and United States are between 5 and 8 percent [Cross *et al*., 2009; Kowalski and Limber, 2007; Sourander *et al*., 2010; Hasebrick, 2011].

Large scale of the problem can be explained with the common access and usage of technology on one hand, since our lives are more and more dependent on mobile solutions. On the other hand – the difficulty in its detection rises from the natural and imposed evolution of the process.

3   State-of-the-Art CB Detection Tools

Till now, a number of researches and solutions have been addressing the matter of CB. These projects included prevention measures like creating games to educate children in the matters of bullying and social exclusion [FearNot!], applications monitoring Twitter for suicidal signals [Samaritans Radar] or analyzing user messages for cyberbylling and informing about the possibility of its harmful content [ReThink].

Many of these solutions are based on regular English words, phrases and constructions, and as far as this approach is relevant – the nature of the Internet imposes tackling the CB problem from other sides, including non-standard lexicons. This paper is first to address the usage of CRS in cyberbullying.

As discussed by Smith [2014], Internet slang and acronyms (Table 1) are used to hide the true nature of online conversations and 1,500 schools across UK using new software to spot dangerous phrases.

Entry:	Explanation:
dirl	"die in real life"; used to upset or encourage to commit suicide
gnoc	"get naked on camera"; used in grooming and sexting
np4np	"naked picture for naked picture"; same as above
IHML	"I hate my life."
oreo	racist slang term for a black person who is 'trying to be white'
mos	"mom over shoulder"
miw	"mom is watching"
pos / pob	"parents over shoulder"
taw	"teachers are watching"
kpc	"keeping parents clueless"

Table 1. Sample Internet slang lexicon

This paper introduces a mixed, two-step approach addressing the discussed issues in online usage of British Cockney Rhyming Slang for cyberbullying and crime-related discussion concealment, namely: slang recognition and conversion for CB and crime detection.

4 Cockney Rhyming Slang

The origins of slang are not certain. Most popular theories consider it being a linguistic game or a cant (cryptolect) among traders / thieves / immigrants of the 19th century London. "Cockney" literally means "a person born within earshot of the bells of St Mary-le-Bow church" (current City of London). Since the bells were destroyed in the WWII bombings in 1945 – the term expanded to not only those from East London but also to London in general.

4.1 Cockney Rhyming Slang Phrasal Construcion

Rhyming slang is based on replacing a common word with a phrase or set of words that rhyme with the replaced word, e.g.: "stairs" is changed with "apples and pears", thus a sentence "*I climbed the stairs*" in Cockney rhyming slang is "*I climbed the apples and pears*".

In Cockney rhyming slang, the lexicon creation appears to have none but one rule – the words have to rhyme. That is why it is common to use non-related words and phrases to replace the keywords, e.g. *becon and eggs = legs*, *air force = sauce*. This lexical latitude has even allowed to replace verbs with nouns ("*Can you Adam and Eve (=believe) it?*"). The usage of names is yet another feature of CRS, including for example British musicians (*Acker Bilk = milk*) and popculture figures (*Ali G = pee*), but also international politicians (*Angela Merkel = circle (roundabout)*) and leaders (*Mahatma Gandhi = shandy*).

4.2 Problems and opportunities in converting CRS

The freedom of the CRS lexicon creation as described above leads to several challenges in creating a tool for transforming CRS into simple English.

First of all, there are many words and phrases in the CRS to replace the same word, e.g. the word "alone" could have been replaced with "Jack Jones" (singer), "Pat Malone" (MLB pitcher), "Todd Sloan" (thoroughbred horse racing jockey) or "Toblerone" (Swiss chocolate bar brand).

Although this situation should not be an issue for a CRS converting tool, since one can imagine creating a list of all slang combinations matching certain words, in some situations rhyming slang uses words that are context and / or gramatically dependent,e.g., "Al Capone" can replace the

noun "telephone" and act as a verb "to be Al Caponed" ("to be stoned").

Code-mixig Resemblance

This problem is a major flaw in one of the biggest assets when working with a CRS convertion, namely its code-mixing resemblance. As defined by Myers-Scotton [1993] "code mixing is the use of words, affixes, phrases and clauses from more than one language within the same sentences". Although slang is not another language *per se*, rhyming slang is based on a word/phrase for word/phrase replacement and it does not affect any other parts and features of the sentence, e.g. tense or word order. This feature narrows the slang "translation" challenge to converting slang sentences and phrases.

5 Lexicon and Corpus Creation

To create CRS lexicon, we have used the *Cockney Rhyming Slang: London's Famous Secret Language* website[1] which currently stores almost 2100 entries.

For the experiment, we have created a corpus made of extracted samples used on the mentioned website[1] and collected all 949 sentences available, out of which 940 were actuall CRS usage examples.

The lexicon entries were manually divided into 3 basic groups. Firstly, we have selected cyberbullying-related words which included: cursing, bad words, insults and words that can be used for sexual harassment and hate speech. Second group consisted of crime-related words: guns and drugs, pedophilia, prostitution, stealing. Finally, other words were marked as neutral.

Category:	Number of entries:
Cyberbullying-related	509
Crime-related	95
Neutral	1430

Table 2. Lexicon categorization statistics

6 Experiments Results

To test the reliability of both modules (recognizing and deciphering; cyberbullying and crime detection) before implementing them into system we have tested both of them as separately.

Baselines and Record Unification

To test the CRS recognition, we have tested raw lexicon entries on the created corpus. The system was able to recognize 519 cases of CRS which set the baseline at 55,43%.

Hemiteleia Problem

The biggest issue to tackle when working with CRS is the so called *hemiteleia* (Ancient Greek, "stop in half") problem. It's the practice of omitting part of the slang term, most often – the rhyming word. For example, earlier discussed sentence "*I climbed the apple and pears*" would actually be abbreviated to "*I climbed the apples*". Understandable at first sight in this example, in other cases hemiteleia can make a sentence impossible to convert, e.g. "She has a nice April" can mean "flowers" or "arse".

To deal with the fact that 382 of the sentences in the corpus (40,64%) were of hemiteleia type, we have modified the system to recognize lexicon format as lists of words and thus – improved the system's precision to 95,85%.

The system was not able to recognize slang in 39 cases. Main reasons of failure were: spelling problems (misspelings and losing the "h" in the beginning of words; 74%), usage of irregular plural forms (13%) and wrongly created entries on the website (13%). For full error analysis, see Appendix 1.

CB and Crime Recognition Experiments

To test the system's CRS detection capabilities, we have selected at random 100 sentences from the corpus containing 108 slang words.

	Overall	CB	crime	neutral
Precision	1.000 (92/92)	1.000 (24/24)	1.000 (7/7)	1.000 (61/61)
Recall	0.851 (92/108)	0.750 (24/32)	0.875 (7/8)	0.873 (61/68)
F1-measure	0.920	0.857	0.9333	0.932

Table 3. CRS detection results

	Overall	CB	crime	neutral
Precision	0.977 (42/43)	1.000 (10/10)	0.750 (3/4)	1.000 (29/29)
Recall	0.388 (42/108)	0.313 (10/32)	0.375 (3/8)	0.426 (29/68)
F1-measure	0.555	0.476	0.500	0.597

Table 4. CRS convertion results

---

[1] http://www.cockneyrhymingslang.co.uk/

7 Conclusions and Future Work

In this paper we have introduced a Cockney rhyming slang recognition system. The system recognizes Cockney Rhyming Slang with an overall F1-score of 0.920 and an abilty to convert sentences containing Cockney slang with an overall F1-score of 0.566.

By high precision in CRS recognition and convertion (1.000 and 9.777 respectively), the system have proven its usefulness, revealing also, how an input-sensitive process working with slang is.

This experience will be taken into consideration while extending the slang lexicon with other British slangs and experimenting with them to improve the system's overall performance in the future and develop the prommising idea behind slang recognition and conversion for cyberbullying and crime detection.

References

[Chen *et al.*, 2005] Haipeng (Allan) Chen, Sharon Ng, Akshay R. Rao. *Cultural Differences in Consumer Impatience*. Journal of Marketing Research, Vol. 42, No. 3, pp. 291-301, August 2005,.

[Cross *et al.*, 2009] Donna Cross, Therese Shaw, Lydia Hearn, Melanie Epstein, Helen Monks, Leanne Lester and Laura Thomas. *Australian Covert Bullying Prevalence Study (ACBPS)*. Child Health Promotion Research Centre, Edith Cowan University, Perth, 2009.

[Cyberbullying Awareness] CyberBullying Awareness. https://sites.google.com/site/cyberbullyingawareness/traditional-bullying-vs-cyberbullying. Access: 04.20.2016.

[FearNot!] Michael Kriegel, Joao Dias, http://sourceforge.net/projects/fearnot/. Access: 04.20.2016.

[Hasebrick *et al.*, 2009] Uwe Hasebrink, Sonia Livingstone Sonia, Leslie Haddon and Kjartan Olafsson. *Comparing children's online opportunities and risks across Europe: Cross-national comparisons for EU Kids Online*. LSE, London: EU Kids Online, 2009.

[Hasebrick 2011] Uwe Hasebrink, Anke Gorzig, Leslie Haddon, Veronika Kalmus, Sonia Livingstone. *Patterns of risk and safety online. In-depth analyses fromthe EU Kids Online survey of 9-16 year olds and their parents in 25 countries*. LSE, London: EU Kids Online, 2011.

[Li, 2007] Li Qing. *Bullying in the New Playground: Research into Cyberbullying and Cyber Victimisation*. Australasian Journal of Education Technology, 23 (4): 435–454, 2007.

[Kann *et al.*, 2014] Laura Kann; Steve Kinchen, Shari L. Shanklin, Katherine H. Flint; Joseph Hawkins, William A. Harris, Richard Lowry, Emily O'Malley Olsen, Tim McManus, David Chyen, Lisa Whittle, Eboni Taylor, Zedwitu Demissie, Nancy Brener, Jemekia Thornton, John Moore,Stephanie Zaza. *Youth Risk Behavior Surveillance — United States, 2013*. Morbidity and Mortality Weekly Report (MMWR), Centers for Disease Control and Prevention, 63(4): 66, 2014.

[Kowalski and Limber 2007] Robin M. Kowalski, Susan P. Limber. *Electronicbullying among middle school students*. Journal of Adolescent Health 41 (6 Suppl 1):S22–S30, 2007.

[Myers-Scotton, 1993] Carol Myers - Scotton. *Social Motivations For Code switching: Evidence from Africa*. Oxford University Press, 1993.

[Pyzalski, 2012] Pyzalski Jacek. *From cyberbullying to electronic aggression: typology of the phenomenon*. Emotional and Behavioural Difficulties, 17: 305-317, 2012.

[ReThink] ReThink, Inc. http://www.rethinkwords.com/. Access: 04.20.2016.

[Samaritans Radar] Samaritans. http://www.samaritans.org/. Access: 04.20.2016.

[Smith 2014] Emily Kent Smith. *From IHML (I hate my life) to Mos (mum over shoulder): Why this guide to cyber-bullying slang may save your child's life*. http://www.dailymail.co.uk/news/article-2673678/Why-guide-cyber-bullying-slang-save-childs-life-From-IHML-I-hate-life-Mos-mum-shoulder.html. Access: 04.20.2016.

[Sourander *et al.*, 2010] Andre Sourander, Anat Brunstein Klomek, Maria Ikonen, Jarna Lindroos, Terhi Luntamo, Merja Koskelainen, Terja Ristkari, Hans Helenius. *Psychosocial riskfactors associated with cyberbullying among adolescents: A population-based study*. Archivesof General Psychiatry, Vol. 67, No. 7, pp. 720–728, 2010.

Appendix 1. Detailed error analysis:

Term extracted:	Meaning extracted:	Source Sentence:	Expected Translation:	Error analysis:
Lah-di-dah	Cigar	Two pints of bitter and a Lahdi please.	Two pints of beer and a cigar please.	spelling
Lanzarote	Tottie	Some Lanzanrote around here!	Some tottie around here!	spelling
Looby Lou's	Shoe	Seen my Loobies?	Seen my shoes?	irregular plural form
Mammy's Smiles	Piles (haemorrhoids)	That vindaloo is playing hell with my Mammies.	That shoe is playing hell with my piles.	irregular plural form
Martini	Lambourghini	Yeah, I'll ave a Lambourghini.	Yeah, I'll ave a Lambourghini.	wrongly created entry
Melody Lingers	Fingers	I had to give up the violin; my Melodies were too short.	I had to give up the violin; my fingers were too short.	irregular plural form
Michael Schumachers	Knackers (testicles)	Interestingly, Schumacher is German for Cobbler too!	Interestingly, Knackers is German for Cobbler too!	spelling
Oats's And Barley	Charlie (cocaine)	Let's have a toot of ya Oats.	Let's have a toot of ya Charlie.	spelling
Porky Pies	Lies	Stop tellin' me Porkies!	Stop tellin' me Lies!	irregular plural form
Potatoes in the Mould	Cold	Blimey it's taters!	Blimey it's Cold!	spelling
Princes Trust	Bust	Look at the Prince's on her.	Look at the Bust on her.	spelling
Prison Time	Bird (Lime)	'Es gawn down, doin' bird.	'Es gawn down, doin' time.	wrongly created entry
Raspberry Ripple	Nipple	Look at the raspberries on that bint.	Look at the nipples on that bint.	irregular plural form
Rolls-Royce	Choice	Its your Rolls.	Its your choice.	spelling
Rub-A-Dub	Pub	Fancy the Rubber?	Fancy the pub?	spelling
Rubik's Cube	Tube	Shall we go on the Rubic's?	Shall we go on the tube?	spelling
Schindlers List	Pissed (drunk)	Down the alehouse, getting Schindler's …	Down the alehouse, getting pissed …	spelling
Sharone Stone	Phone	No worries china give us buzz on the Sharon.	No worries mate give us buzz on the phone.	spelling
Sooty And Sweep	Sleep	I could do with some Soot.	I could do with some sleep.	spelling
Stinging Nettle	Kettle	Put the stinger on.	Put the kettle on.	spelling

Term extracted:	Meaning extracted:	Source Sentence:	Expected Translation:	Error analysis:
Ten Ounce Rump	Dump	I'm going for a 10er.	I'm going for a dump.	spelling
Toblerone	Alone	I've been sitting here for hours on my Tobler.	I've been sitting here for hours on my alone.	spelling
Becks and Posh	Money (Dosh)	I can't tonight, I've got no Beck's.	I can't tonight, I've got no money (dosh).	spelling
Beggar boy's arse	Brass (old Cockney for a prostitute)	She's a bit brassy.	She's a bit brass.	wrongly created entry
Blaine	Insane	Jeez mate - that's a bit David innit?	Jeez mate - that's a bit insane innit?	wrongly created entry
Arthur Conan Doyle	Boil	Clive, get that kettle on the arfur!	Clive, get that kettle on the boil!	spelling
Hampstead Heath	Teeth	Got to go down the dentist to have me 'Ampsteads looked at.	Got to go down the dentist to have me teeth looked at.	spelling
Harris Tweed	Weed	He's a bit of an 'Arris.	He's a bit of an weed.	spelling
Harry Flint	Skint	I'm a bit 'Arry at the moment.	I'm a bit skint at the moment.	spelling
Harry Potter	Snotter (nose)	That geezer's got an 'arry like concorde!	That geezer's got an nose like concorde!	spelling
Henry the Eighth	Eighth (of cannabis)	Sort me out an Enry.	Sort me out an eighth.	spelling
Jay Kay	Take Away	Fancy a jay-kay after this?	Fancy a take away after this?	spelling
Jellied Eel	Squeal	Darren you f**kin' Jelly.	Darren you f**kin' squeal.	spelling
Jellied Eel	Deal	I'll give ya a great jelly on it.	I'll give ya a great deal on it.	spelling
Calvin Klein	Fine (nice body)	Oi Oi you're Clvin Kein!	Oi Oi you're fine!	spelling
Cobbler's Awls	Balls	That's a load of cobblers, for a start!	That's a load of balls, for a start!	spelling
Coochie-Coo	Zoo	We're going to the coochie coo today.	We're going to the zoo today.	spelling
Crabs	Doner Kebabs	That old brass gave me the Doner Kebabs.	That old brass gave me the crabs.	wrongly created entry
Dawson's Creek	Streak	I remember Erica doing a Dawsons like it was yesterday.	I remember Erica doing a streak like it was yesterday.	spelling