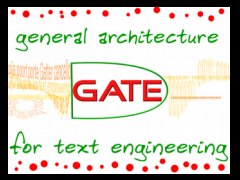
Handling and Mining Linguistic Variation in UGC

Leon Derczynski



What is UGC?

What is UGC?

No editor

Comes direct from end user


NYSE
NYSE
NYSE
NYSE
NYSE
NYSE
NYS
CYPRESS
PERFORM


NYSE
NYSE
NYSE
NYSE
NYSE
NYSE
NYS
CYPRESS
FREEDOM


NYSE
CYPRESS


NYSE
NYSE
NYSE
NYSE
NYSE
CYPRESS

Streaming social media is powerful

- It's Big Data!
  - Velocity: 500M tweets / day
  - Volume: 20M users / month
  - Variety: earthquakes, stocks, this guy

- Sample of all human discourse - unprecedented
- Not only where people are & when, but also *what* they are doing
- Interesting stuff - just ask the NSA!

Tweets are *dirty*

- You all know what Twitter is, so let's just look at some difficult tweets
- Orthography: `Kk its 22:48 friday nyt :D really tired so imma go to sleep :) good nyt x god bles xxxxx`
- Fragments: `Bonfire tonite. All are welcome, joe included`
- Capitalisation: `Don't Have Time To Stop In??? Then, Check Out Our Quick Full Service Drive Thru Window :)`
- Nonverbal acts: `RT @Huddy85: @Mz_Twilightxxx *kisses your ass**sneezes after* Lol`

Tough tweets: Do we even care?

- Most tweets are linguistically fairly well-formed
- `RT @DesignerDepot: Minimalist Web Design: When Less is More - http://ow.ly/2FwyX`
- `just went on an unfollowing spree... there's no point` *`of`* `following you if you haven't tweeted in 10+ days. #justsaying ..`
- The tweets we find most difficult, are those that *seem* to say the least
- `So im in tha chi whts popping tonight?`
- `i just gave my momma some money 4 a bill.... she smiled when i put it n her hand __AND__ said "i wanna go out to eat"... -______- HELLA SCAN`

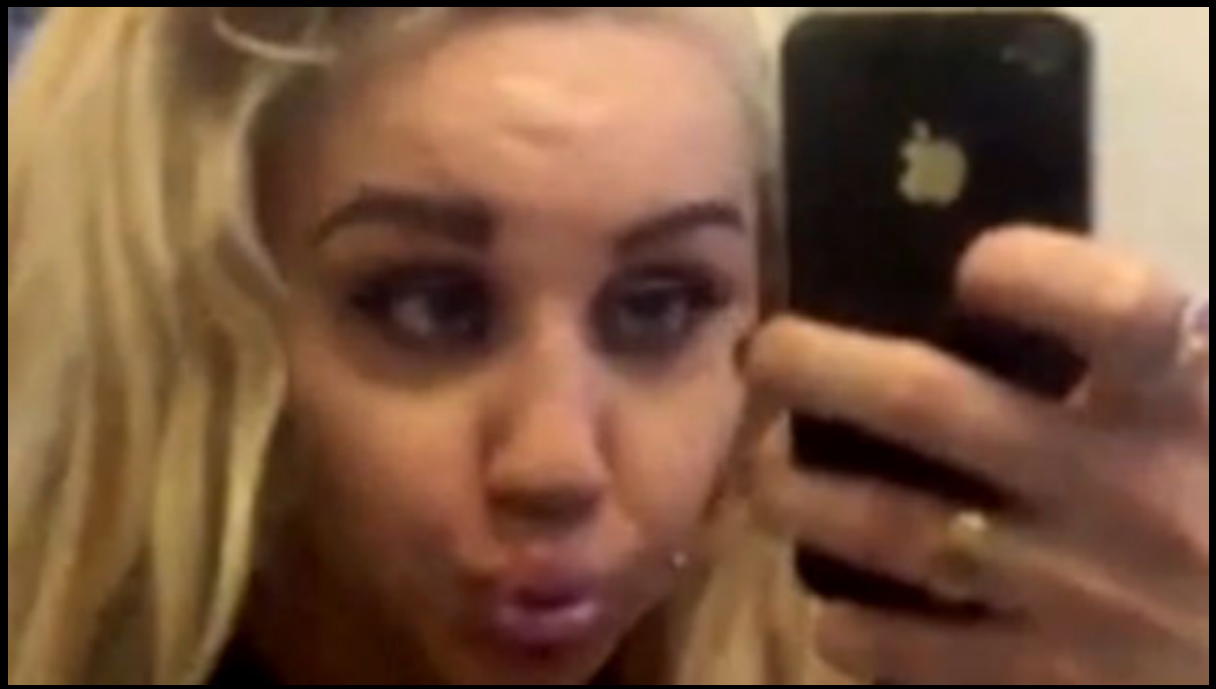
Word order still matters.. just

- Hard for tweets: exclamations and fragments
- Whole sequences a bit rare
- `@FeeninforPretty making something to eat, aint ate all day`
- `Peace green tea time!! Happyzone!!!! :)))))`

- Sentence structure cues (e.g. caps) often:
  - absent
  - over-used

How much variation is there?

social media text is **surprisingly** formal



they see me rollin

- a typo?

they see me rollin
they hatin

- perhaps not. G-dropping mapped from speech

they see me rollin
they hatin
patrollin

they see me rollin
they hatin
patrollin
tryna catch me ridin dirty

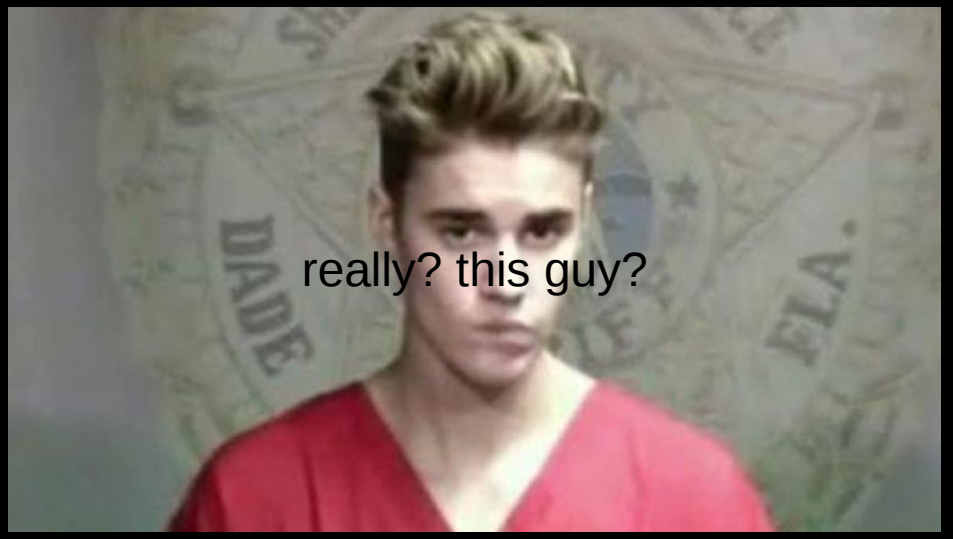
- flawless; not a single mistake

omb **x**

- surely they mean “omg”*?*

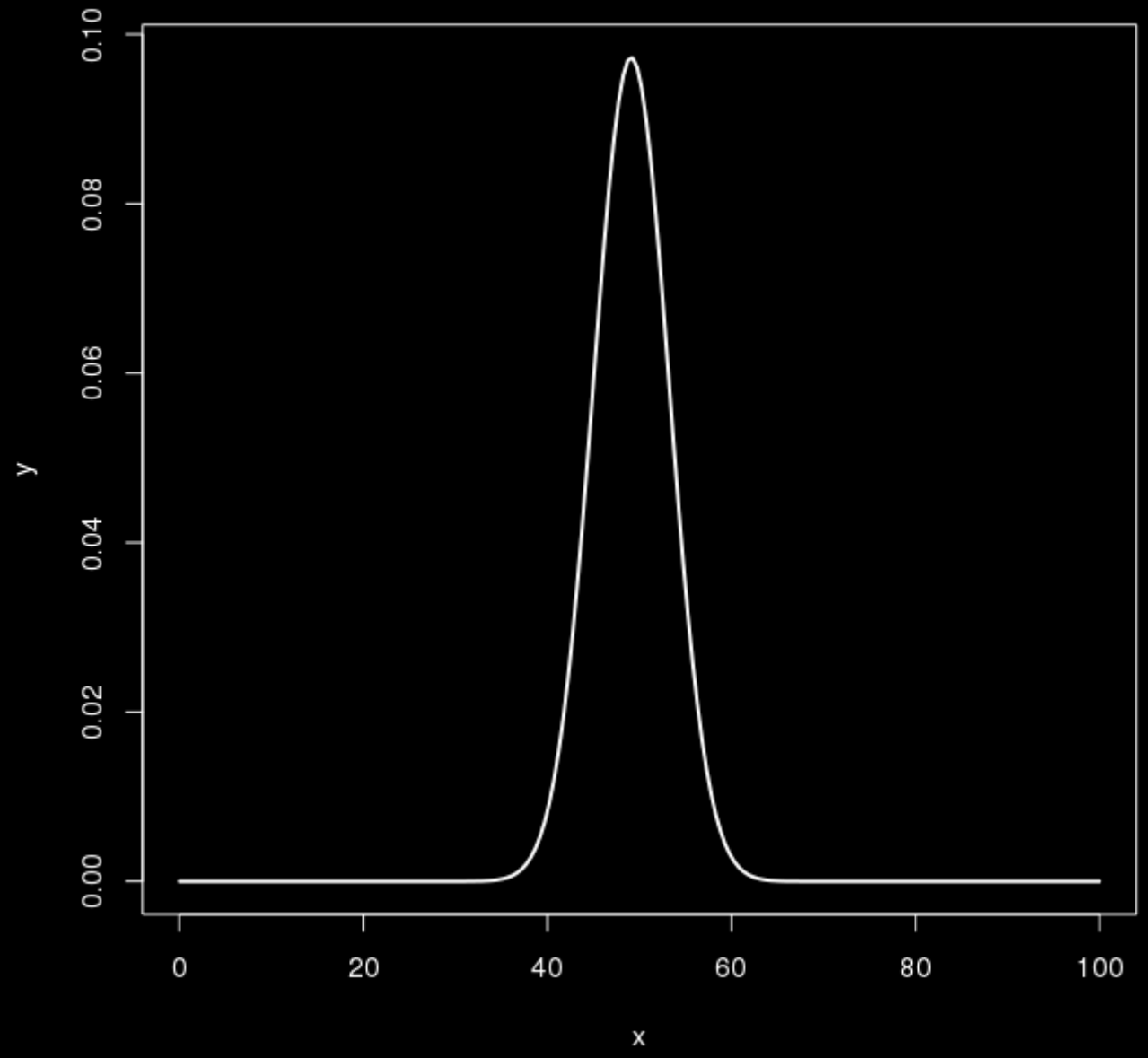
omb ✔

- the keys are like, right next to each other

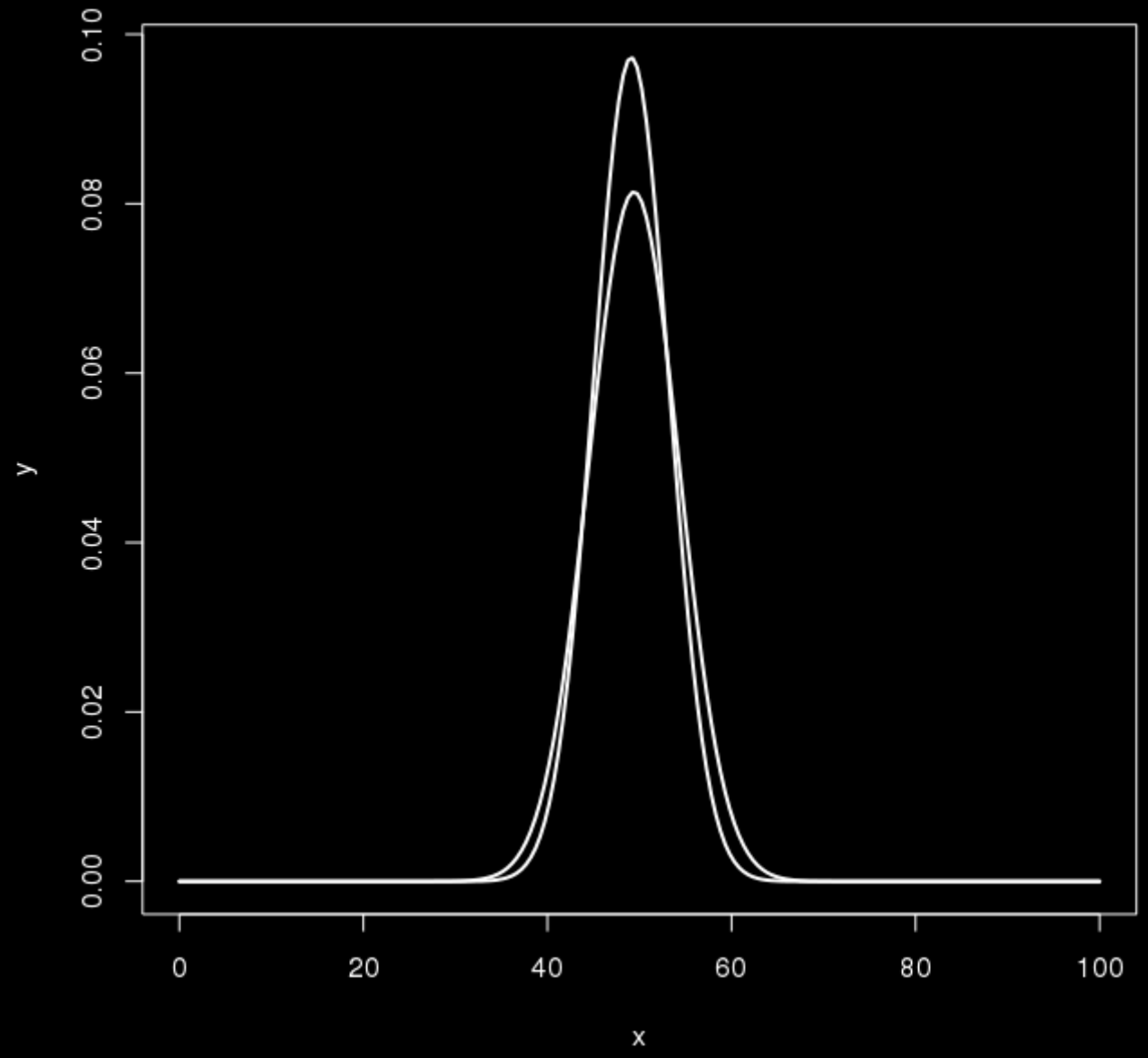


Shall we go out for dinner this evening?

Ey yo wen u gon let me tap dat



spelling ability distribution in net slang users



with spelling ability distribution in non-slang users

Drouin 2009, J. Literacy Research

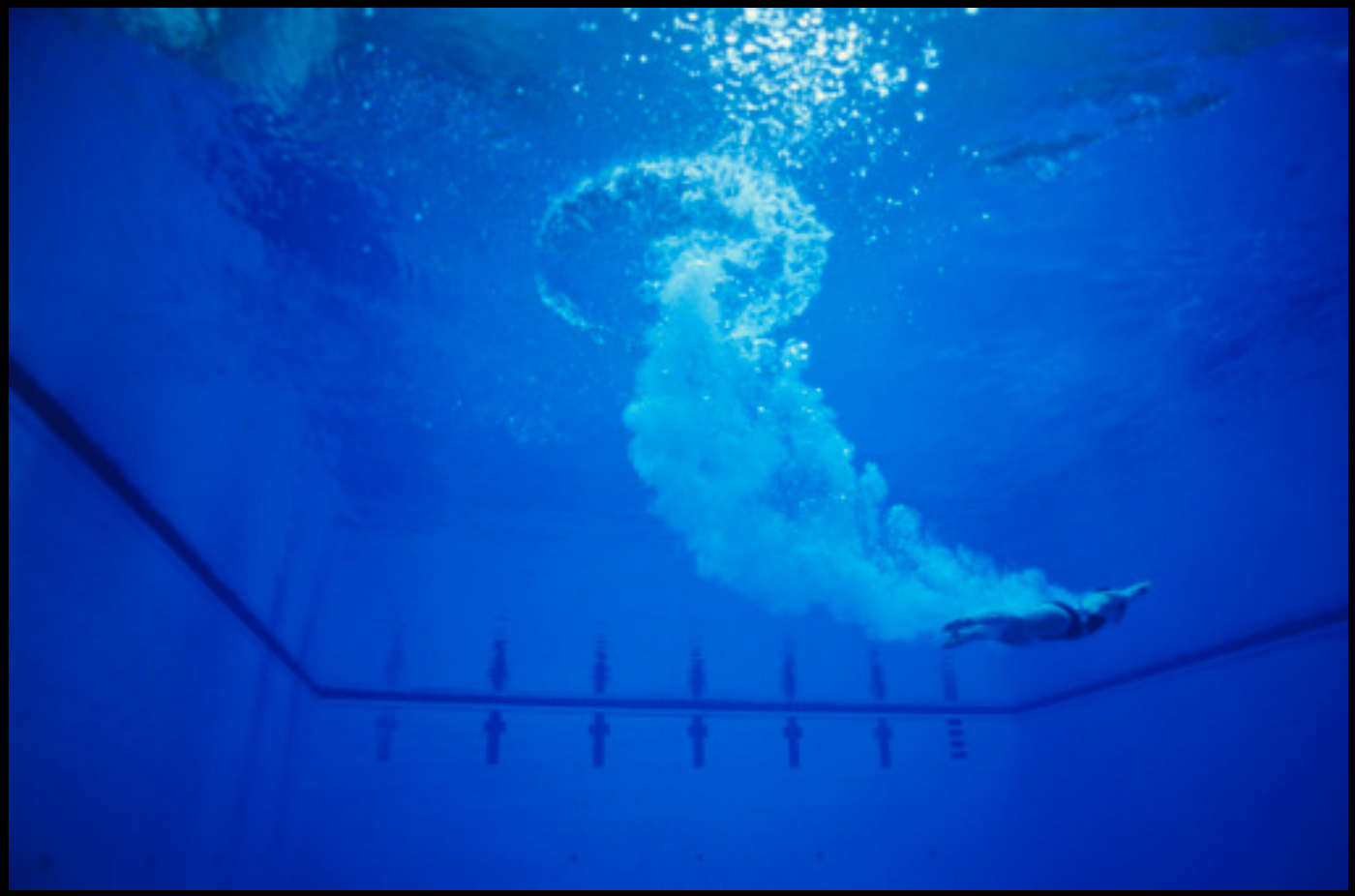
General challenges

- Common complaints we have about social media text:
  - Documents are short;
  - There are spelling errors;
  - Words are ambiguous;
  - Nonstandard / new lexical items;
  - Nonstandard syntactic patterns.

- The impact (or the cause?) of these complaints: Low performance of existing systems.
  - Maybe we need to re-train?
  - Shortage of training data;
  - Low-performance of existing techniques.

General challenges

- How can we characterise UGC?
- What form does the variation take?



Qualitative description

- Great diversity in social media users, but they're not illiterate
  - People want to represent their own dialects and accents (Jones 2010)
  - They pick and choose from the entire stylistic repertoire of language (Tagliamonte 2008)
  - Same literacy scores in standard and non-standard vocabulary users (Drouin 2009)
- Emoticons have more than just an expressive function
  - Pragmatic function, e.g. demonstrating a less stressed stance (Dresner 2010)
  - Not just pictograms: phrasal abbreviations are also included – smh, lol
  - Lexical items are made nonstandard through lengthening – cooolll (Brody 2011)

Qualitative description

- Social variables associated with certain transformations
  - Slang is less inhibited in informal settings (Labov 1972)
  - G-dropping mapped from speech to writing (Eisenstein 2010)
  - Lexemes can have a spatial association within a language (Eisenstein 2011)
- This socio-linguistic variation in social media highlights bias in existing resources
  - Most corpora text was curated predominantly by working-age white men (Eisenstein 2013)
  - Social media is not curated, so has different biases
  - We have little data that is free from this demographic bias

Quantitative description

- General style
  - Twitter is more conservative and formal, less conversational than SMS and online chat;
  - Brief, with careful word choice, dense in lexical words (Halliday 2004);
  - Tweets used to share news or broadcast personal status
- Individual style:
  - Individualistic style, or address large audience? (Yates 1996)
  - Users develop linguistically unique styles
  - Both 1st and 3rd person pronouns are common
  - Intensifiers of a young audience - “really” vs. “very” (Ito 2003).

What errors occur on unknowns?

- Gold standard errors (`dank_UH je_UH → _FW`)
- Training lacks IV words (`Internet`, `bake`)
- Pre-taggables (URLs, mentions, retweets)
- NN vs. NNP (`derek_NN`, `Bed_NNP`)
- Slang (`LUVZ`, `HELLA`, `2night`)
- Genre-specific (`unfollowing`)
- Tokenisation errors (`ass**sneezes`)
- Orthographic (`suprising`)

Tweet non-standard language

- From analysis, three big issues identified:

1. **Many unseen words / orthographies**

2. **Uncertain sentence structure**

3. **Not enough annotated data**

- Continued with Ritter dataset

Unseen words in tweets

- Two classes:
- Standard token, non-standard orthography;
  - freinds
  - KHAAAANNNNNNN!

- Non-standard token, standard orthography
  - omg + bieber = omb
  - Huntington

challenge 1: what language is this anyway

je bent **Jacques cousteau** niet die een nieuwe soort heeft ontdekt, het is duidelijk, ze bedekken hun gezicht. **Get over it**

RT @TomPIngram: VIVA LAS VEGAS 16 - NEWS #constantcontact http://t.co/VrFzZaa7

challenge 2: pls type better

I wonde rif Tsubasa is okay..

- misplaced space = two new words

no homwork tonight.. suprising??

- maybe there should be!

challenge 3: finding names

derek

x *is a person*

miles

x *might be a person*

Marie Claire

x *should not be a person*

Exodus Porter

x *probably an OK person, but actually a beer*

challenge 3: finding names

Spicy Pickle Jr.

x *apparently actually a person*

challenge 3: finding names

Spicy Pickle Jr.

x *apparently actually a person*

???

How can we process UGC?

Distributional representations

- “Lexical variation creates sparsity and OOV”
- Model words using Brown clustering (classes + tree)
- Only a bigram model – from 1992!

- Cluster for “*tomorrow*”: *(Ritter et al., EMNLP'22)*
  - 2m, 2ma, 2mar, 2mara, 2maro, 2marrow, 2mor, 2mora, 2moro, 2morow, 2morr, 2morro, 2morrow, 2moz, 2mr, 2mro, 2mrrw, 2mrw, 2mw, tmmrw, tmo, tmoro, tmorrow, tmoz, tmr, tmro, tmrow, tmrrow, tmrrw, tmrw, tmrww, tmw, tomaro, tomarow, tomarro, tomarrow, tomm, tommarow, tommarrow, tommoro, tommorow, tommorrow, tommorw, tommrow, tomo, tomolo, tomoro, tomorow, tomorro, tomorrow, tomorrw, tomoz, tomrw, tomz
- Other embeddings are fine too, though clusterless
- Generate the right # of classes

  (Derczynski et al., RANLP'15)

Distributional representations

- Map new-seen words into Brown clusters
- Use FST to find Brown clusters with similar patterns

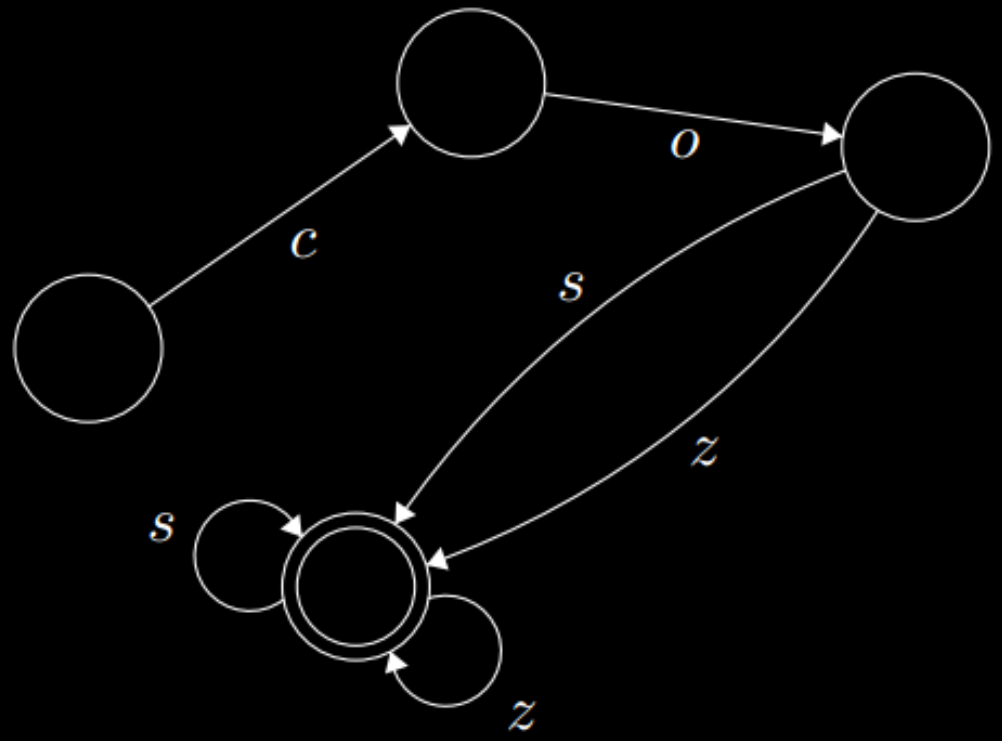


Figure 1: Example FSA for Brown cluster {cos, coz, coss, cozzz}

Wulff & Søgaard, WNUT '15

Unseen words in tweets

- Majority of non-standard orthographies can be corrected with a gazetteer: typical Pareto
  - `vids → videos`
  - `cussin → cursing`
  - `hella → very`

- No need to bother with e.g. Brown clustering
- 361 entries give 2.3%/token PoS error reduction

Distributional representations

- Normalisation
  - Convert twitter text to “well-formed” text; e.g. slang resolution
  - Some success using noisy channel model (Han 2011)
  - Techniques include: edit distance; double metaphone with threshold
  - Issues: false positives can change meanings, e.g. reversing sentiment (apolitical; unarmoured)
- Domain adaptation
  - Treat twitter as its own domain / genre, and create customised tools and techniques
  - Some success in language ID (Carter 2013), PoS tagging (Gimpel 2011), NER (Ritter 2011)
  - Millions of micro-domains…. that's more than one!

Unseen words in tweets

- The rest can handled reasonably with word shape and contextual features
- Using edu.stanford.nlp.tagger.maxent.ExtractorFramesRare
- Features include:
  - word prefix and suffix shapes
  - distribution of shape in corpus
  - shapes of neighbouring words
- Sadly, shape often useless..
  - last day of sorting pope visit to birmingham stuff out
  - Don't Have Time To Stop In??? Then, Check Out Our Quick Full Service Drive Thru Window :)

Capitalisation

- Noisy tweets have unusual capitalisation, right?
  - `Buy Our Widgets Now`
  - `ugh I haet u all .. stupd ppl #fml`
- Lowercase model with lowercased data allows us to ignore capitalisation noise
- Tried multiple approaches to classifying noisy vs. well-formed capitalisation
- Gain from ignoring case in noisy tweets offset by loss from mis-classified well-cased data

Leveraging author information

- Demographics help!
  - Hovy 'ACL 14
  - Trained embeddings over whole corpus, and also sub-corpora segmented demographically
  - Segmented representations performed better at topic sentiment, topic, demographic classification

- Hmm, interesting..

What does the variation tell us?

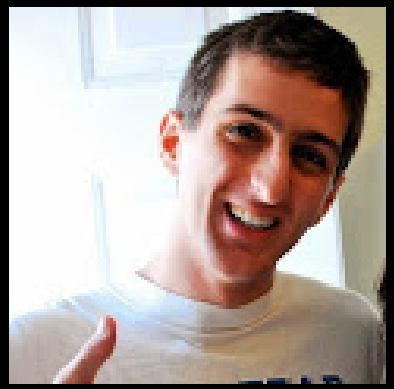
Do you feel luccy, punk?

Do you feel luccy, punk?

“Now we stronger than ever”

- Let's look at AAVE
- Hyp 1:
  - Ethnicity represented in syntax
- Hyp 2:
  - Greater specialised syntax use by females

- We already have a dataset…



Ian Stewart, EACL 2014

“Now we stronger than ever”

- AAVE already affects our toolchain
- Twitter tokeniser (CMU)
  - Slang contractions segmented differently “ima”
- Part-of-speech tagging
  - Nonstandard tense-aspect-mood not in tagsets
    - “finna”
  - Need to detect 3rd person sing. “s” drop
    - “oh man she run”
  - Use both CMU and PTB tagset

“Now we stronger than ever”

Construction	Example from Corpus	Simplified Pattern	Tagger Used
copula deletion	we stronger than ever	not(V)+PRO+ADJ	PTB
habitual *be*	now i be sober af	not(V)+PRO+*be*+ADJ	PTB
continuative *steady*	steady getting bigger	*steady*+not(N)	Gimpel
completive *done*	u done pissed me off	*done*+V_{PST}	PTB
future *finna (fixing to)*	i’m finna tweet	*finna*+V	Gimpel
remote past *been*	i been had it	PRO/N+*been*+V_{PST}	PTB
negative concord	don’t say nothing	*don’t/ain’t/can’t*+V+ *nobody/nothing/nowhere/no*	Gimpel
null genitive marking	time of they life	PRO_{NOM}+N	Gimpel
ass camouflage construction (Collins et al. 2008)	divorced his ass	V+PRO_{POSS}+*ass*	PTB

“Now we stronger than ever”

- Correlate ling var with ZCTA: regional demographics
- AAVE more common in urban areas
  - Standard American English prevails in rural
- Positive match when all variations are combined
  - Not so strong for individual traits
- Tendency to support for gender hypothesis
  - Though not with negation syntax!

AAVE in Canonical context

- False positives:
  - Copula deletion in Singapore English
  - Quoted text, e.g. rap lyrics – is this a persona?
- “finna” presents all kinds of problems
  - Should be VP, not NN
  - Systems to handle non-standard morphemes don't exist

User reviews for sociolinguistics

- Start with reviews from Trustpilot
  - Danish, Dutch, English, Finnish, French, German, Italian, Norwegian, Polish, Portuguese, Russian, Spanish, Swedish
- Users are labelled with location, age, gender

- Variation in language then
  - Mined (χ^2 then aggressive *p*-test)
  - Used to predict each of these three categories

Hovy et al. WWW 2015

User reviews for sociolinguistics

- Location
- Informal and Formal variations

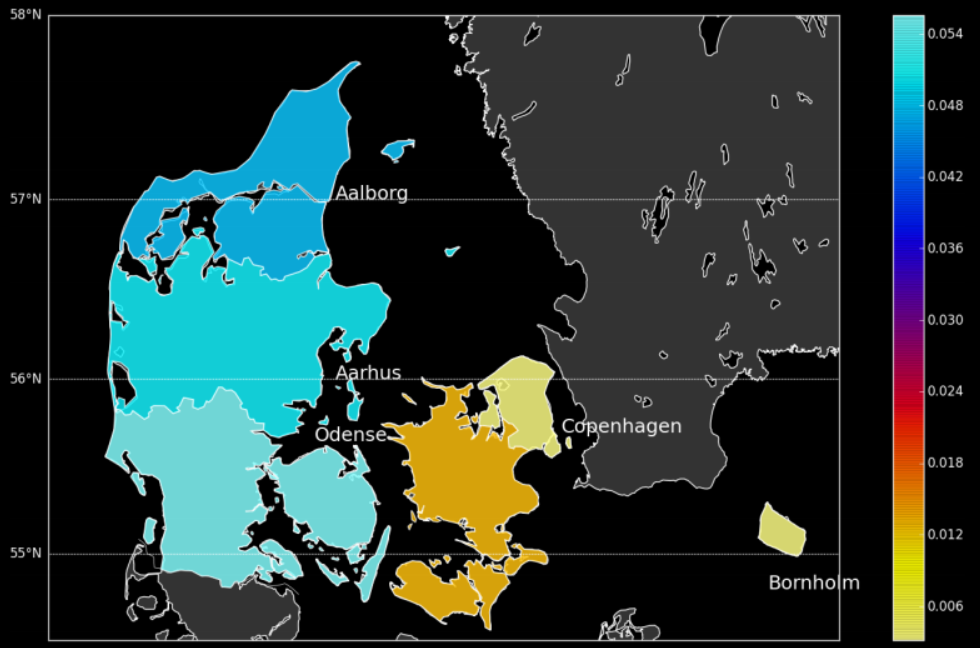


Figure 6: Regional distribution of the word *trælls* (ratio of all words per NUTS-2 region).

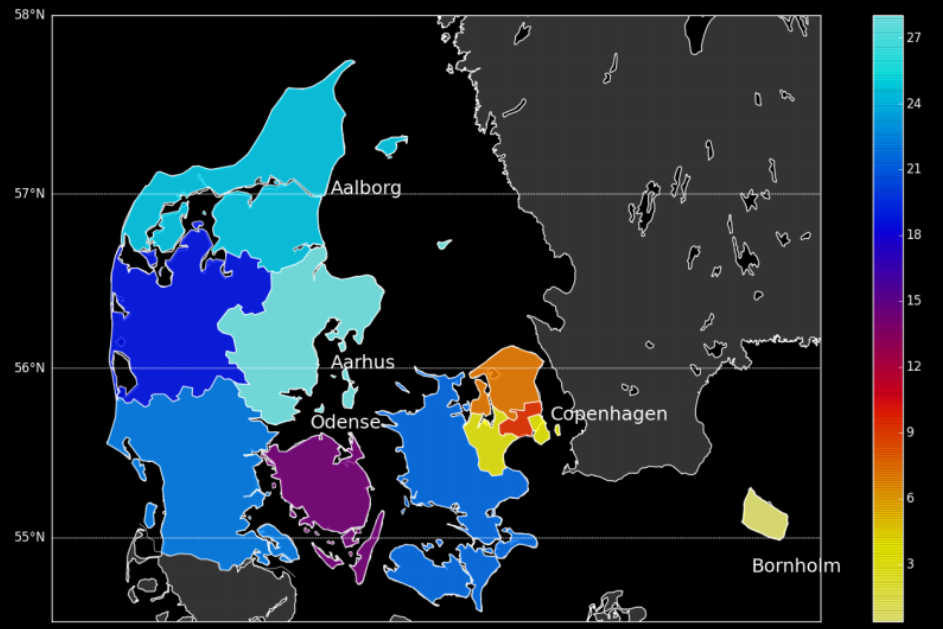


Figure 5: Regional distribution of "sin/sit" (ratio of all pronouns per NUTS-2 region)

User reviews for sociolinguistics

- Gender

	Denmark	France	Germany	UK	US
female	nemt	**satisfaite**	gerne	flowers	customer_service
	bestille	**contente**	qualität	really	**love**
	varer	suis	tolle	**pleased**	we
	tilfreds	produits	wir	thank_you	**pleased**
	rigtig	chaussures	ware	definitely	really
	handle	étiquettes	bestellen	lovely	petflow
	dejligt	sont	sind	**love**	free_shipping
	nem	ma	karten	quickly	thank_you
	hjemmeside	été	bestellung	received	our
	mine	spartoo	vielen_dank	impressed	food
male	lovet	**problème**	keys	wife	these_guys
	deres	rien	günstig	parts	parts_geek
	problemer	bonne	service	deal	fast
	alt	matériel	mmoga	product	partsgeek
	aftalt	**conforme**	support	00	prices
	virker	ldlc	weiter_so	an	best
	tiden	service	top	first_class	shipping
	ok	sans	key	tyres	customer
	pris	bon	guter	could_find	part
	00	satisfait	gerne_wieder	best	parts

User reviews for sociolinguistics

- Gender

- Females use emoticons twice as often
- Top product categories:
  - F: Pets, clothes/fashion, hotels, fashion accessories
  - M: Computers, car lights, hotels, fashion accessories
- Females *slightly* prefer noseless smileys
- Same distribution of review scores

User reviews for sociolinguistics

- Age - Biggest indicator: NOSES

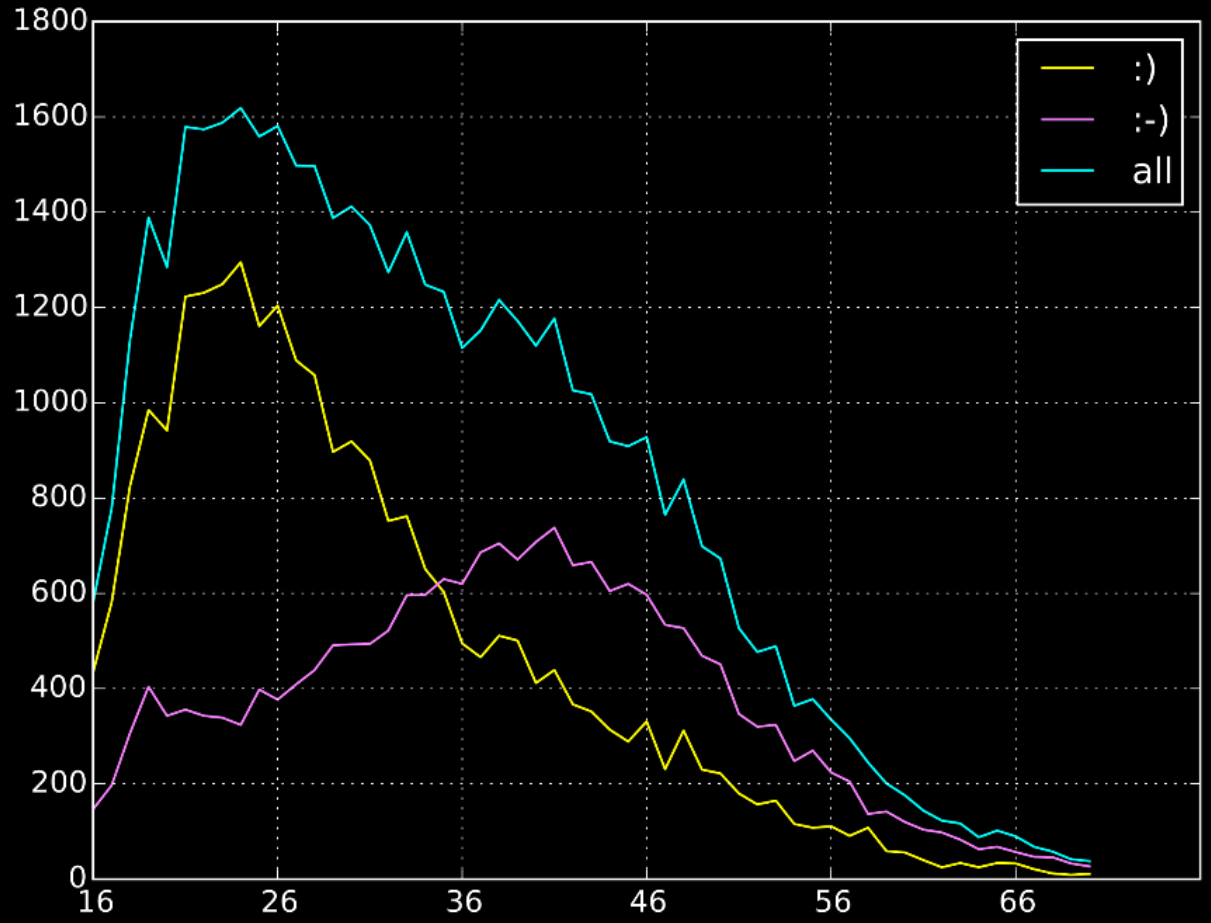
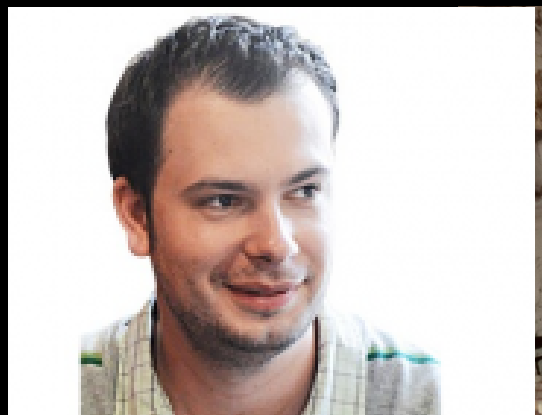


Figure 3: Usage of emoticons with and without nose by age group, aggregated over all countries

Use a nose?
You're 35+!

User reviews for sociolinguistics

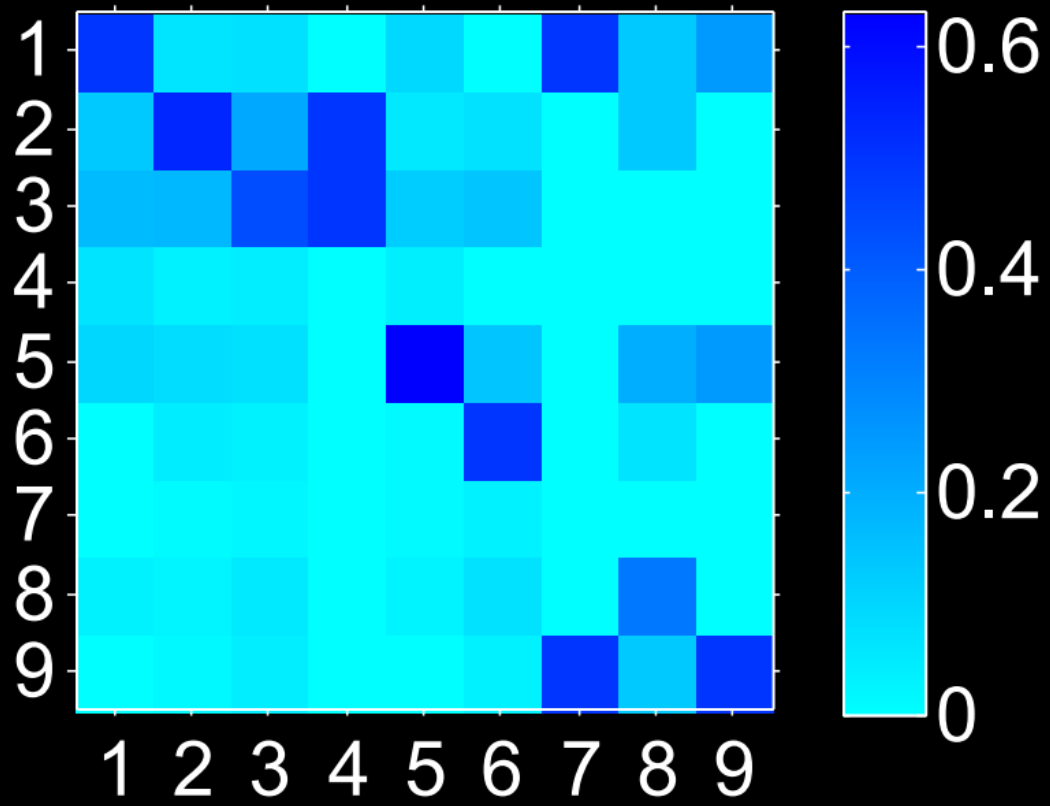
- Job *(Preotiuc, Lampos, Aletras ACL'15)*
- 9-class system: Management, professional, skilled trades, sales etc
- Data from Twitter:
  - Followers & friends
  - Hashtag, RT, link usage
  - Tweet rate
  - % in English



User reviews for sociolinguistics

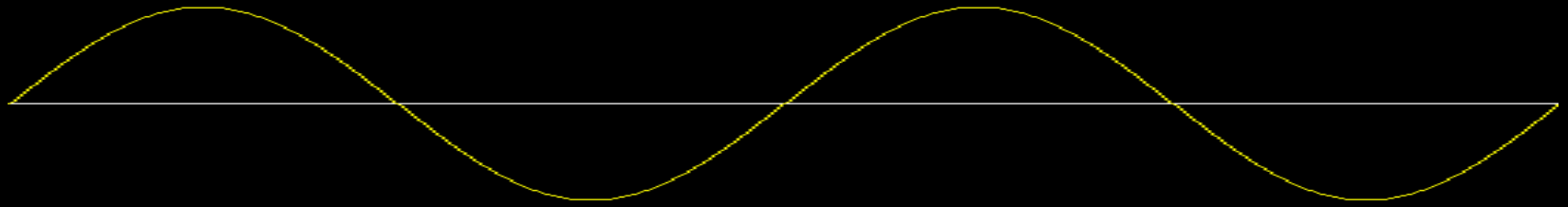
- W2V and SVD feature extraction
- Clusters better than embeddings

- Reasonable accuracy
- Enables mining of:
  - Related topics
  - Behaviour



Where to go from here

news



social media

most of our language AI was trained on news text

the bias is:

- middle class
- white
-working age
- educated
- male
- 1980s/1990s
- from the US
- journalist
- following AP guidelines

your phone rewards you if you talk and write like

(ok.. sort of)

your phone rewards you if you talk and write like

(ok.. sort of)

.. and punishes you when you don't.

(not cool!)

the REAL problem:

our tools have been built around a
tiny, over-biased set of data

there is no variation!
(annotate some WSJ if you are not convinced..)

it's time to up our game,
and the field is noticing the utility in variation

e.g. Baldwin @ WNUT15; Hovy @ ACL15

Variation tells us a whole amount
– including how to process itself

Let's capture real-world variation from UGC,
and use it to do NLP for the real world

Thank you for your time!

There is hope:

`Jersey Shore is overrated. studying and history homework then a fat night of sleep!`

Do you have any questions?