## Sentiment Analysis

НИС «Компьютерная семантика» Даша Рыжова, Даша Попова

### В формальной семантике:

- Independence: Expressive content contributes a dimension of meaning that is separate from the regular descriptive content.
- 2. Nondisplaceability: Expressives predicate something of the utterance situation.
- Perspective dependence: Expressive content is evaluated from a particular perspective. In general, the perspective is the speaker's, but there can be deviations if conditions are right.
- Descriptive ineffability: Speakers are never fully satisfied when they paraphrase expressive content using descriptive, i.e., nonexpressive, terms.
- Immediacy: Like performatives, expressives achieve their intended act simply by being uttered; they do not offer content so much as inflict it.
- Repeatability: If a speaker repeatedly uses an expressive item, the effect is generally
  one of strengthening the emotive content, rather than one of redundancy.

### В формальной семантике:

(45)a. damn:  $\langle e, \varepsilon \rangle$ b.  $[\![\mathbf{damn}]\!]^c = \text{the function } f \text{ such that } f([\![a]\!]^c)(c) = c', \text{ where }$ i.  $c \approx_{c_{I}, \llbracket a \rrbracket^{c}}^{\mathbf{I}} c';$ the length of **I** is not more than .5; and iii.  $\mathbf{I} \subseteq [-1, 0]$ (47)a. **fucking**:  $\langle e, \varepsilon \rangle$ **[fucking]**  $^c$  = the function f such that  $f([a]^c)(c) = c'$ , b. i.  $c \approx_{c_I, \llbracket a \rrbracket^c}^{\mathbf{I}} c';$ ii. the length of I is not more than .2; and iii.  $\mathbf{I} \subseteq [-1, 0]$ Potts, 2007

### Экспрессивы и восклицания

Constant et al., 2009; Potts and Schwarz, 2008

Rating: 1 star

Review: I bought this book and read it half way through. It was soo damn boring to read. Just

opinions on what caused what. It was absolutely dreadful and was a good way to fall

asleep. Sold it back on amazon for 8 bux.

(a) A short one-star review

Rating: 3 star

Review: This is a good book to compare the various menu items and products for their calorie

values. The one problem I have with it is that it only shows 3-4 good items and 3-4 bad items per restaurant/store/manufacturer. I would rather have had this broken into a

series of books and gone into more detail.

(b) A short three-star review

Rating: 5 star

Review: I never had an opinion either way about Tori. I thought she was fun in Scary Movie 2,

and then got totally addicted to So NoTORIous. I couldn't believe how down to earth and self-deprecating she was, and it seemed so sincere. Her book confirmed that too! And I learned that in the end, little girls (rich or poor)are the same, and all we really

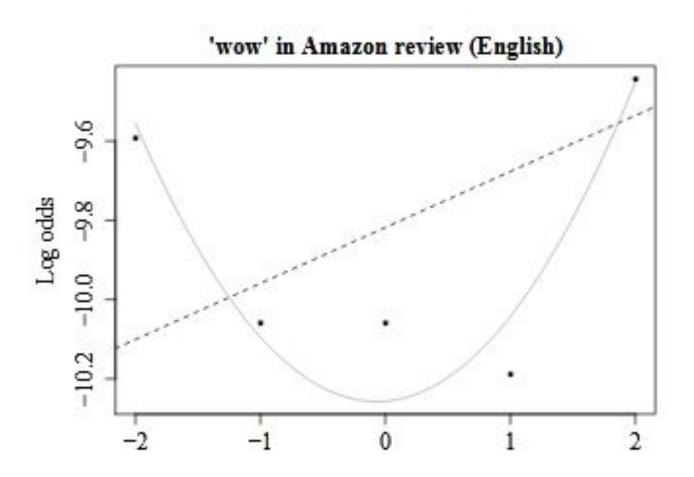
want is that damn Barbie dream house :-)

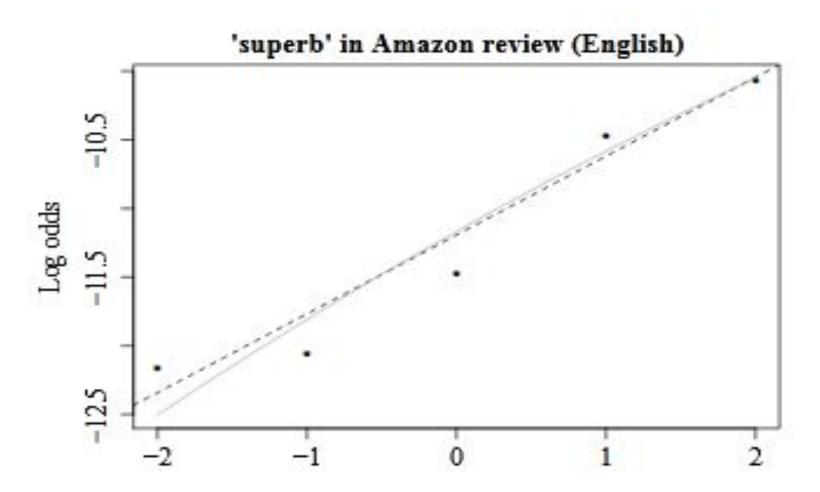
(c) A short five-star review

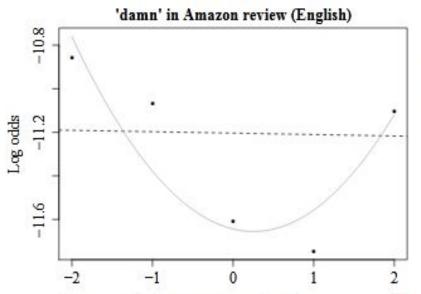
Figure 1: Example reviews from Amazon.com

Our basic perspective is the log-odds distribution, as in (1), where  $count(x_n, R)$  is the number of tokens of  $x_n$  (a word-string of length n) in rating category R, and  $count_n(R)$  is the number of tokens of word-strings of length n in rating category R.

(1) 
$$\log\text{-odds}(x_n, R) \stackrel{def}{=} \ln \left( \frac{\text{count}(x_n, R)}{\text{count}_n(R) - \text{count}(x_n, R)} \right)$$

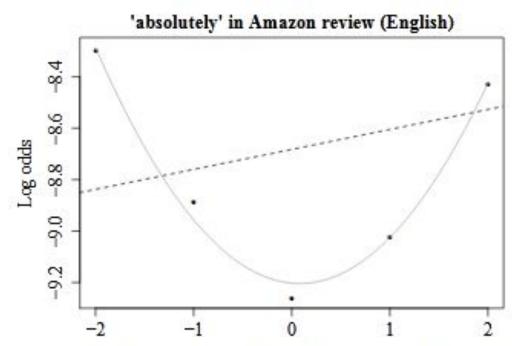






- a. Try answering the damn question.
- b. Sounds like another damn politician to me.
- c. Just a damn minute! What history books did you read?

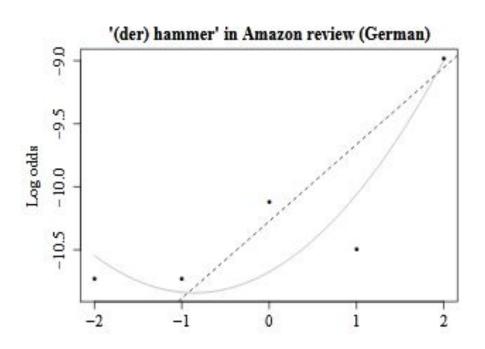
- a. funnest damn movie i've seen all year<sup>6</sup>
- b. Linnea Faris, a woman from Michigan who was wearing a "Remember Alex" T-shirt, shook her head in disbelief. [...] "I've spent hours crying over that damn bird."



- a. I could absolutely/totally/SO jump over that fence.
- b. MTV like totally gave us TWO episodes back to back. It was like so random. The more the merrier, but it's like waay too much for one recap. (Web example from Beaver and Clark 2008: 74)

### Экспрессивы: немецкий

- a. Dieses Album ist der Hammer
   This album is the hammer
   'This album is cool.' (≈ 'the bomb', colloquially)
- amys stimme mischt sich mit hammer beats
   Amy's voice mixed self with cool beats
   'Amy's voice mixed with cool beats.'
- c. Was f
  ür ein Hammer Album what for a cool album 'What a cool album!'

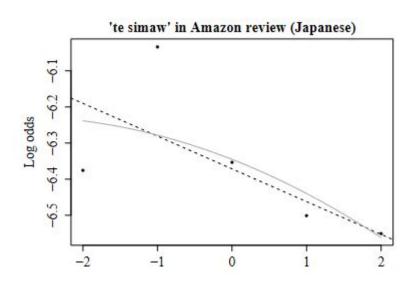


## Экспрессивы: японский

(11) harii pottaa to wa nagai o-tukiai de, owat-**tesimat**-te totemo kanasii Harry Potter with торіс long ноn-acquaintance be end-**ANTIHON**-INF very sad omoi-ga si-masu thought-nom do-ноn

'I had a long acquaintance with the Harry Potter series, and was sad to see it end.'

(12) nesugoshi-chimat-ta.
overslept-antihon-past
'I overslept, which sucks.'



### Восклицания в.

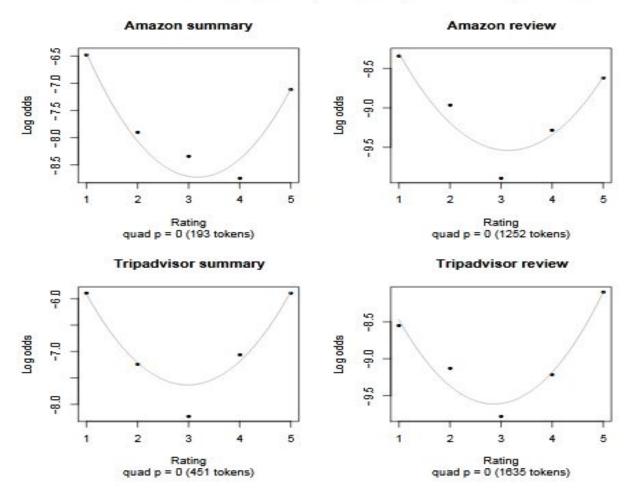
- a. What a Find!
- . What an awesome place!!
- c. What a mess!
- d. What an Overpriced Dump

(from a five-star Tripadvisor review)

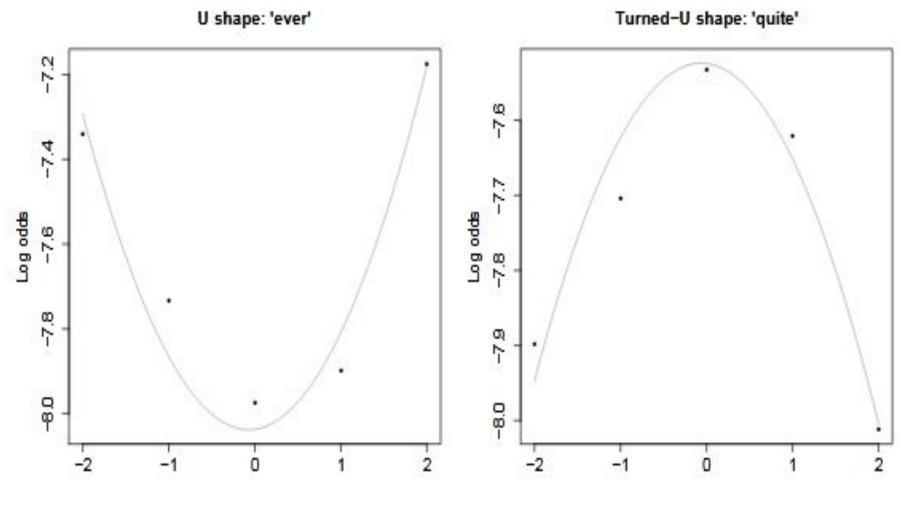
(from a five-star Tripadvisor review)

(from a one-star Tripadvisor review)

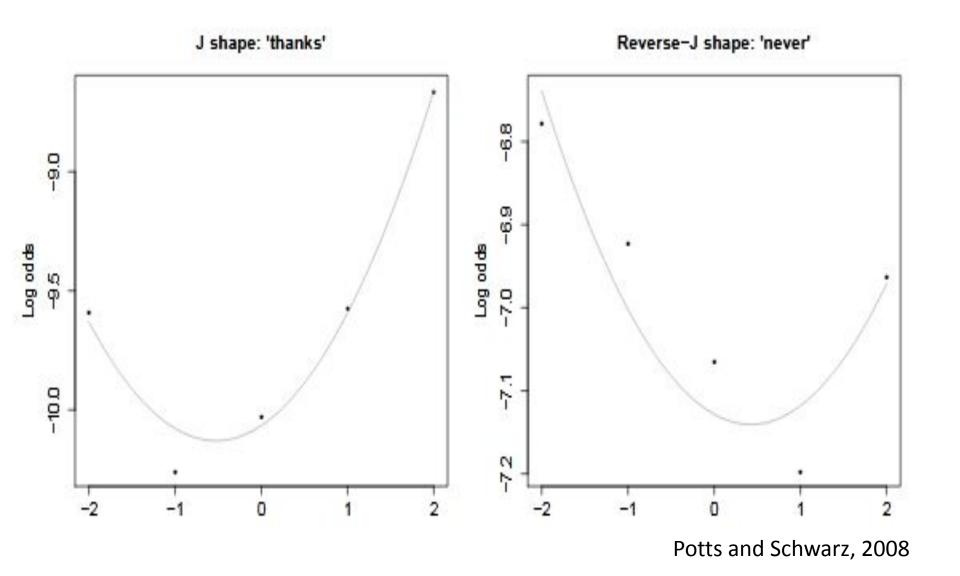
(from a one-star Tripadvisor review)



Potts and Schwarz, 2008



Potts and Schwarz, 2008



11	!	absolutely	all
best	ever	every	i've ever
i've	i	it!	my
the best	this	what a	wow

(a) The items whose shapes are limited to U, J, and Reverse J for all four corpora.

! !	!	absolute	absolutely	
again!	all	am	any	
anyone	best	book	couldn't	
even	ever!	ever	ever had	
every	have ever	i	i am	
i could	i have	i've	i've ever	
is the	it!	life	must	
my	never	new	one of	
simply	the best	this	this is	
time	what	what a	will	
will never	wow!	wow		

<sup>(</sup>b) The items whose shapes are limited to U, J, and Reverse J in at least three of the four corpora.

- but	a few	a good	average
basic	but nothing	but some	but still fine
cons	decent	few	
little too	mostly	near the	not bad
not quite	otherwise	part	points
pretty	pros	quite	short
some with some	somewhat	though	two

<sup>(</sup>c) The items whose shapes are limited to Turned U in at least three of the four corpora.

# Экспрессивные указательные местоимения

## До Трампа: Sarah Palin

"Americans are cravin' that straight talk"

## Бурная реакция на ее речь

#### FoxNews.com:

- "We feel like she talks like we do."
- "She talked like real people to real people"

#### **Huffington Post:**

- "illusion of straight-talking"
- "pseudo-folksiness and fundamental dishonesty"

## Дебаты 2008

Joe Biden, Palin's opponent:

"We should be helping them build schools to compete for those hearts and minds of the people in the region"

## И все, все, все:

- В полиции: make that phone call right now
- Стюард(ecca): get those bags under that seat in front of you
- Инструктор йоги: get that left arm up over that head

# Экспрессивные употребления указательных местоимений

- давно известны лингвистам
- есть и в русском:
  - Смех смехом, а в полицию уже обратилось более двадцати горожан, которые уверяют, что после встречи с этим "похитителем человеческих душ" (так окрестили маньяка журналисты) у них чего-то не хватает. Ох, и мнительные эти американцы! У нас бы на такого маньяка народ молился. Лос-анджелесская полиция злоумышленника усиленно разыскивает, но пока безуспешно. [Похититель душ // «Криминальная хроника», 2003.06.24]
  - эти дети мне уже надоели
  - Уберите уже эти ноги из прохода!
  - на вкус отвратительно— ну и гадость эта ваша заливная рыба [Наши дети: Подростки (2004)]
  - Какого полёта птица этот ваш человекомух и какой он там подвиг совершил— нам с Жоржем плевать. [Олег Павлов. Карагандинские девятины, или Повесть последних дней // «Октябрь», 2001]
  - Что вы мне талдычите, мог он порисоваться в так называемом узком кругу, что вы мусолите мочалу... Этот ваш прогрессивный реформатор в двадцать девятом году барду для самогона добывал, тем его связь с сельским хозяйством и ограничивалась. [Анатолий Азольский. Лопушок // «Новый Мир», 1998]

## Корпусный подход

Эти утверждения импрессионистические

Можем ли мы их квантифицировать?

Например, проанализировать дистрибуцию экспрессивных местоимений по аннотированным контекстам?



```
Confession: I bought a case of beer, now I'm watching a South Park
            marathon while getting drunk :P
 Reactions: Sorry, hugs: 2; You rock: 3; Teehee: 2, I understand: 3;
             Wow, just wow: 0
Confession: subconsciously, I constantly narrate my own life in my head. in
            third person, in a british accent. Insane? Probably
 Reactions: Sorry, hugs: 0; You rock: 7; Teehee: 8; I understand: 0;
             Wow, just wow: 1
Confession: I really hate being shy ... I just want to be able to talk to some-
            one about anything and everything and be myself... That's all
            I've ever wanted.
 Reactions: Sorry, hugs: 1; You rock: 1; Teehee: 2; I understand: 10;
             Wow, just wow: 0;
```

#### 10 Reactions











Figure: EP reaction icons.

Sorry, hugs sympathy

You rock cheering, supportive

Teehee amused

I understand solitary

Wow, just wow shock

Table: Interpreting the icons.

		Category	Reactions
		Sorry, hugs	91,222 (22%)
		You rock	80,798 (19%)
Texts	140,467	Teehee	59,597 (14%)
Words	21,518,718	I understand	125,026 (30%)
Vocab	143,712	Wow, just wow	60,952 (15%)
Mean words/text	153.19	Total	417,595
	X		

<sup>(</sup>a) The overall size of the corpus.

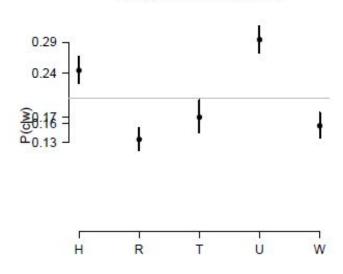
(b) All reactions.

Table: In general, reader reactions are sympathetic and supportive.

## Считаем и изображаем

Cat.	Count	Total	$Pr_{EP}(w r)$	$Pr_{EP}(r w)$
Sorry, hugs	1167	18038374	0.00006	0.26
You rock	520	14066087	0.00004	0.15
Teehee	300	8167037	0.00004	0.15
I understand	1488	20466744	0.00007	0.29
Wow, just wow	473	12550603	0.00004	0.15

disappointed - 1,070 tokens



## Указательные местоимения

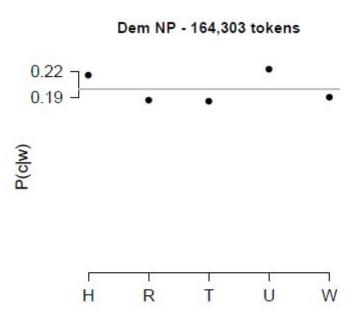


Figure: All determiner demonstratives in the EP data.

### Примеры: слова, вызывающие симпатию

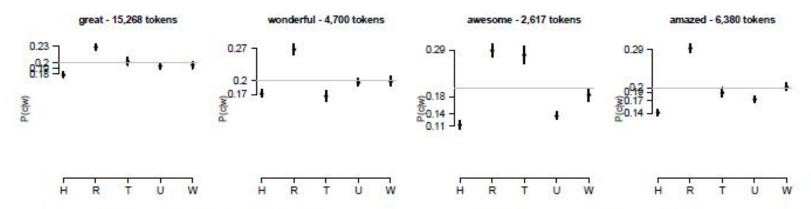


Figure: Words eliciting predominantly 'you rock' reactions.

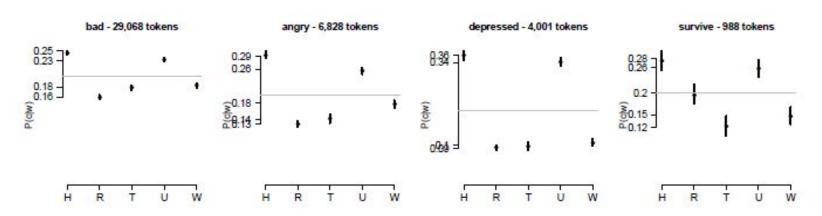


Figure: Words eliciting sympathetic reactions.

- Что же особенного в употреблении указательных местоимений у Palin?
- 16 интервью Palin на шоу
- + интервью до и после нее
- всего 48 интервью

## Квантитативный анализ

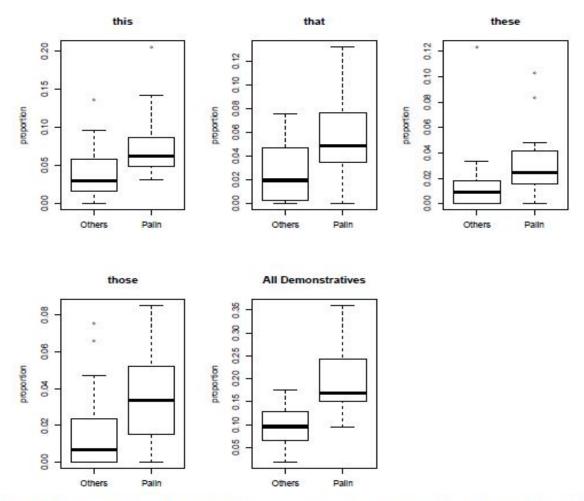


Figure: Non-pronominal dems. as a proportion of all determiners.

## Полярные мнения – почему?

#### FoxNews.com:

- "We feel like she talks like we do."
- "She talked like real people to real people"

#### **Huffington Post:**

- "illusion of straight-talking"
- "pseudo-folksiness and fundamental dishonesty"

## Свойскость

Указательные местоимения в непрямом значении создают эффект **солидарности, свойскости** (термин восходит к лекциям Сандро Васильевича Кодзасова, мой перевод для solidarity из анализа статьи)

Проявления свойскости уместны при совпадении отношения к сообщаемому, мнений

При несовпадении эффект обратный: навязанной солидарности, навязанной близости мнений, отношения к сообщаемому

## Квантитативный анализ

- 1. Группы с указательными местоимениями у Palin длиннее, более описательны, чем у других
- 2. Более провокационный контекст

"these good, hard-working, average, everyday, patriotic Americans who want to see the positive change in our country that they deserve"

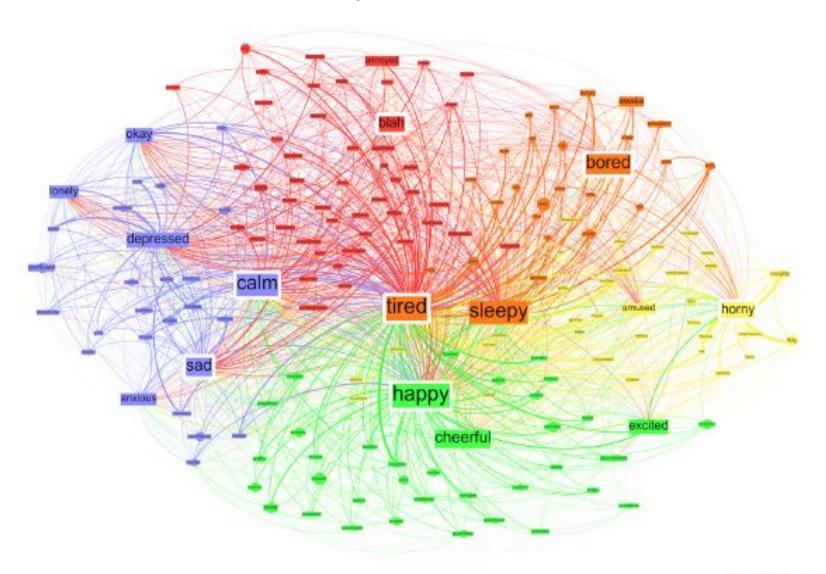
# Sentiment analysis

# Концептуальные проблемы

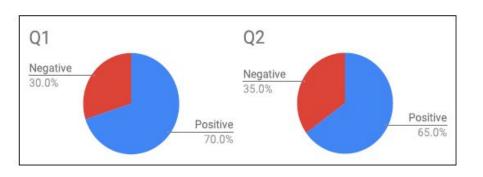
Какие из следующих высказываний выражают отношение (sentiment)? Какова полярность отношения (sentiment polarity: pos/neg), если оно выражено? Достаточно ли нам градации pos/neg?

- 1. There was an earthquake in California.
- 2. The team failed to complete the physical challenge. (We win/lose!)
- 3. They said it would be great.
- 4. They said it would be great, and they were right.
- 5. They said it would be great, and they were wrong.
- 6. The party fat-cats are sipping their expensive imported wines.
- 7. Oh, you're terrible!
- 8. Её муж какой-то сантехник.
- 9. Отличное кино!
- 10. Опять надо выгуливать эту чертову собаку!
- 11. Of 2001, "Many consider the masterpiece bewildering, boring, slow-moving or annoying, . . . "
- 12. long-suffering fans, bittersweet memories, hilariously embarrassing moments, . . .

# Измерения аффекта (valence=polarity; arousal = intensity), отношения между элементами, и переходы



#### Много потенциальных применений, но какова реальная цель?



- Картинки слева не помогают в принятии решений
- Даже если распределение отражает реальность, оно скрывает понастоящему релевантные категории

# Связанные задачи

<ul> <li>Subjectivity</li> </ul>		(Pang and Lee 2008	3)
<ul><li>Bias</li></ul>	(Recasens et	al. 2013; Pryzant et al. 2020	))
<ul> <li>Stance</li> </ul>		(Anand et al. 2011	L)
<ul> <li>Hate-speech</li> </ul>		(Nobata et al. 2016	5)
<ul> <li>Microaggressi</li> </ul>	ons	(Breitfeller et al. 2019	))
<ul> <li>Condescensio</li> </ul>	n	(Wang and Potts 2019	))
<ul> <li>Sarcasm</li> </ul>		(Khodak et al. 2017	7)
<ul> <li>Deception and</li> </ul>	d betrayal	(Niculae et al. 2015	5)
<ul> <li>Online trolls</li> </ul>		(Cheng et al. 2017	7)
<ul> <li>Polarization</li> </ul>		(Gentzkow et al. 2019	))
<ul> <li>Politeness</li> </ul>	(Danes	cu-Niculescu-Mizil et al. 2013	3)
<ul> <li>Linguistic alig</li> </ul>	nment	(Doyle et al. 2016	5)

# Некоторые релевантные • IMDb movie reviews (50K) (Maas et al. ДЕЙТАСЕТЫ:

http://ai.stanford.edu/~amaas/data/sentiment/index.html

• Datasets from Lillian Lee's group: http://www.cs.cornell.edu/home/llee/data/

 Datasets from Bing Liu's group: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

- RateBeer (McAuley et al. 2012; McAuley and Leskovec 2013): http://snap.stanford.edu/data/web-RateBeer.html
- Amazon Customer Review data: https://s3.amazonaws.com/amazon-reviews-pds/readme.html
- Amazon Product Data (McAuley et al. 2015; He and McAuley 2016): http://imcaulev.ucsd.edu/data/amazon/
- Sentiment and social networks together (West et al. 2014) http://infolab.stanford.edu/~west1/TACL2014/
- Stanford Sentiment Treebank (SST; Socher et al. 2013) https://nlp.stanford.edu/sentiment/ https://nlp.stanford.edu/~socherr/EMNLP2013 RNTN.pdf

### Лексика

- Bing Liu's Opinion Lexicon: nltk.corpus.opinion\_lexicon
- SentiWordNet: nltk.corpus.sentiwordnet
- MPQA subjectivity lexicon: <a href="http://mpqa.cs.pitt.edu">http://mpqa.cs.pitt.edu</a>

- Hamilton et al. (2016): SocialSent
   https://nlp.stanford.edu/projects/socialsent/
- Brysbaert et al. (2014): Norms of valence, arousal, and dominance for 13,915 English lemmas
- Sentiment and Emotion Lexicon

http://saifmohammad.com/WebPages/lexicons.html

#### Токенизация

@CompSem: can't wait for the Jun 13 #projects! YAAAAAAY!!! >:-D

```
• Токенизируем по
  пробелам
@CompSem:
can't
wait
for
the
Jun
13
#projects
YAAAAAAY!!!
>:-D
```

#### Токенизация

@CompSem: can't wait for the Jun 13 #projects! YAAAAAAY!!! >:-D

- Токенизируем по пробелам @CompSem: can't

wait

for

the

Jun

23

#projects

YAAAAAY!!!

>:-D

- Элементы токенайзера, полезного для выявления отношения:
- выделяет эмодзи
- определяет разметку твиттера и т.п.
- использует теги типа <strong>
- отлавливает замаскированные #\$%-ми ругательства
- оставляет заглавные буквы там, где это осмысленно
- нормализует удлинения (например,  $YAAAAAAY \rightarrow YAAAY$ )
- отлавливает важные составные выражения (например, out of this world)

nltk.tokenize.casual.TweetTokenizer

## Опасности стемминга

• Теряем информацию о форме слова

Positiv	Negativ	Porter stemmed
defense	defensive	defens
extravagance	extravagant	extravag
affection	affectation	affect
competence	compete	compet
impetus	impetuous	impetu
objective	objection	object
temperance	temper	temper
tolerant	tolerable	toler

## Опасности стемминга

• Теряем информацию о форме слова

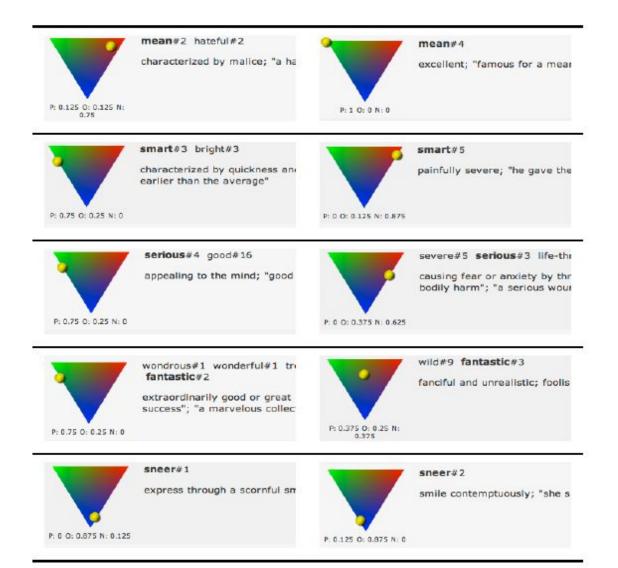
Positiv	Negativ	Lancaster stemmed
call	callous	cal
compliment	complicate	comply
dependability	dependent	depend
famous	famished	fam
fill	filth	fil
flourish	floor	flo
notoriety	notorious	not
passionate	passe	pass
savings	savage	sav
truth	truant	tru

## Опасности стемминга

• WordNet stemmer отлично справляется, кроме сравнительной степени

Positiv	WordNet stemmed	
(exclaims, v)	exclaim	
(exclaimed, v)	exclaim	
(exclaiming, v)	exclaim	
(exclamation, n)	exclamation	
(proved, v)	prove	
(proven, v)	prove	
(proven, a)	proven	
(happy, a)	happy	
(happier, a)	happy	
(happiest, a)	happy	

# Опасности разметки по частям речи Пара (слово, часть речи) может нести как позитивное, так и негативное отношение на уровне леммы



# Опасности отрицания

- 1. I didn't enjoy it.
- 2. I never enjoy it.
- 3. No one enjoys it.
- 4. I have yet to enjoy it.
- 5. I don't think I will enjoy it.

# Опасности отрицания

- 1. I didn't enjoy it.
- 2. I never enjoy it.
- 3. No one enjoys it.
- 4. I have yet to enjoy it.
- 5. I don't think I will enjoy it.

Метод, предложенный в (Das and Chen 2001; Pang et al. 2002):

Append a \_NEG suffix to every word appearing between a negation and a clause-level punctuation mark.

# Опасности отрицания

No one enjoys it.	no one_NEG enjoys_NEG it_NEG
I don't think I will enjoy it, but I might.	i don't think_NEG i_NEG will_NEG enjoy_NEG it_NEG
	but i might

#### Литература 1

- Acton, Eric K. and Christopher Potts. 2014. <u>That straight talk: Sarah Palin and the sociolinguistics of demonstratives</u>. *Journal of Sociolinguistics* 18(1): 3-31.
- Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011.
   Cats rule and dogs drool!: Classifying stance in online debate. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pages 1–9, Portland, Oregon.
   Association for Computational Linguistics.
- Luke Breitfeller, Emily Ahn, Aldrian Obaja Muis, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics.
- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.
- Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '17, pages 1217–1230, New York, NY, USA. ACM.
- Noah Constant; Christopher Davis; Christopher Potts; and Florian Schwarz. 2009. The pragmatics of expressive content: Evidence from large corpora. Sprache und Datenverarbeitung 33(1-2):5-21.
- Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013.
   A computational approach to politeness with application to social factors. In *Proceedings of the 2013* Annual Conference of the Association for Computational Linguistics, pages 250–259, Stroudsburg, PA.
   Association for Computational Linguistics.
- Gabriel Doyle, Dan Yurovsky, and Michael C. Frank. 2016. A robust framework for estimating linguistic alignment in twitter conversations. In *Proceedings of the 25th International World Wide Web Conference, WWW '16, pages 637–648,* Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.
- William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Austin, Texas. Association for Computational Linguistics.

#### Литература 2

- Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2019. Measuring group differences in high-dimensional choices:
   Method and application to congressional speech. Ms, Stanford University, Brown University and Amazon.
- Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2017. A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579*.
- Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic harbingers of betrayal: A case study on an online strategy game. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1650–1659, Beijing, China. Association for Computational Linguistics.
- Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web, WWW '16, pages 145–153,* Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.
- Christopher Potts. 2007. <u>The expressive dimension</u>. *Theoretical Linguistics* 33(2):165-197.
- Christopher Potts [joint research with Eric Acton]. 2013. <u>Cravin' that straight talk: the latent affective meaning of</u> demonstratives. Workshop on Computational Social Sciences, Stanford, Jan 11.
- Christopher Potts and Florian Schwarz. 2008. <u>Exclamatives and heightened emotion: Extracting pragmatic generalizations from large corpora</u>. Ms., UMass Amherst.
- Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1):1–135.
- Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically
  neutralizing subjective bias in text. In *Proceedings of AAAI*.
- Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.
- Moritz Sudhof, Andrés Gómez Emilsson, Andrew L. Maas, and Christopher Potts. 2014. Sentiment expression conditioned by affective transitions and social forces. In *Proceedings of 20th Conference on Knowledge Discovery and Data Mining, pages* 1136–1145, New York. ACM.
- Zijian Wang and Christopher Potts. 2019. TalkDown: A corpus for condescension detection in context. In *Proceedings of the* 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3702–3710, Stroudsburg, PA. Association for Computational Linguistics.