

Natural Language Processing

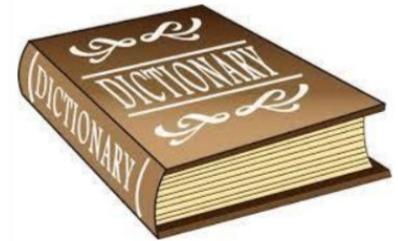
Linguistics Background, Classification

Yulia Tsvetkov

yuliats@cs.washington.edu

Linguistic Background

What does it mean to “know” a language?



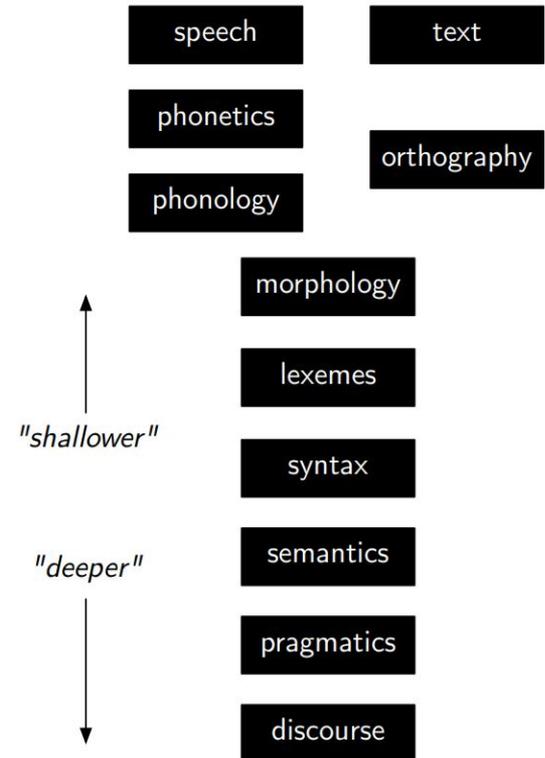
Hi, how can I help?

What do we need to “tell” a computer program so that it knows more English than w_c or a dictionary, maybe even as much as a three-year-old, for example?

What does an NLP system need to 'know'?

- Language consists of many levels of structure
- Humans fluently integrate all of these in producing/understanding language
- Ideally, so would a computer!

Levels of linguistic knowledge

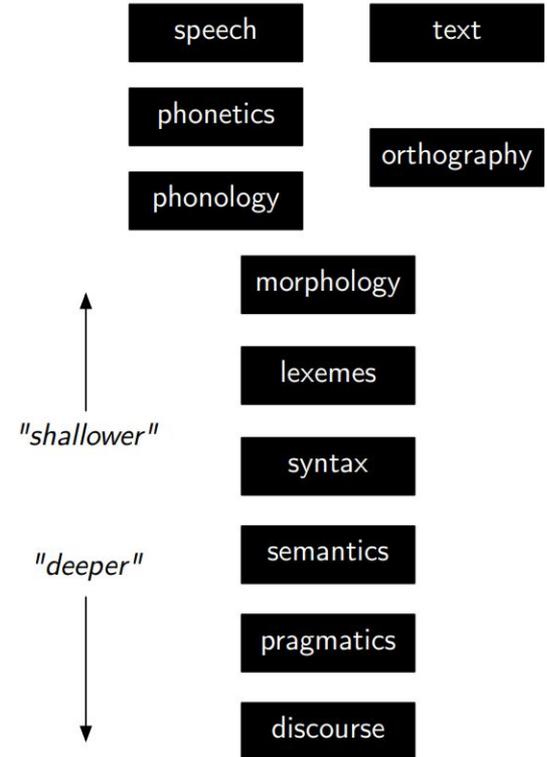


Speech, phonetics, phonology



This is a simple sentence .

/ ðɪs ɪz ə 'sɪmpl 'sɛntəns /.



Orthography

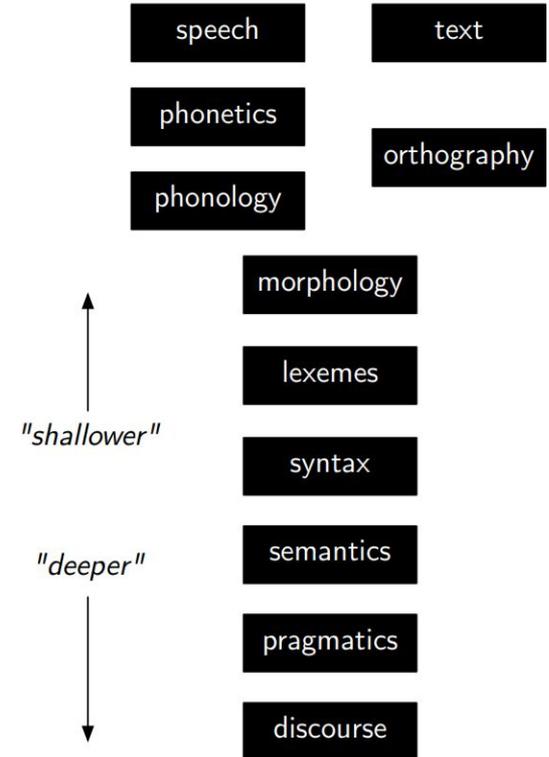
هذه جملة بسيطة

đây là một câu đơn giản

यह एक साधारण वाक्य है

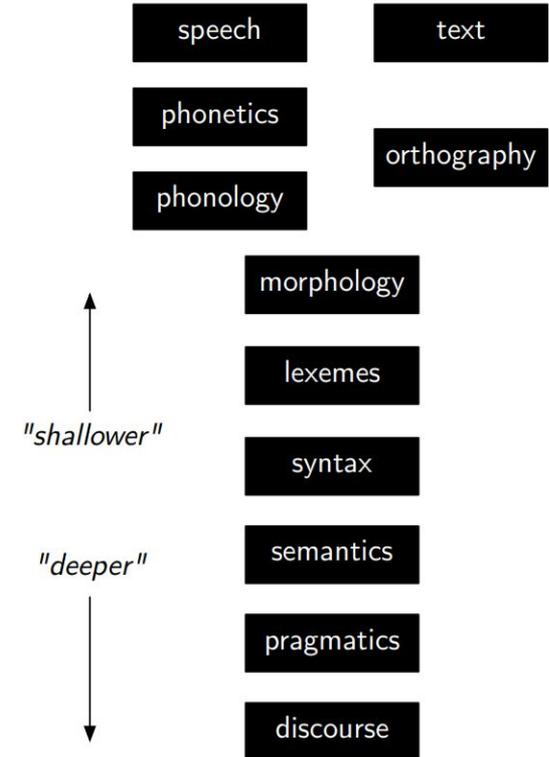
This is a simple sentence .

/ ðɪs ɪz ə 'sɪmpl 'sɛntəns /.



Words, morphology

- Morphological analysis
- Tokenization
- Lemmatization



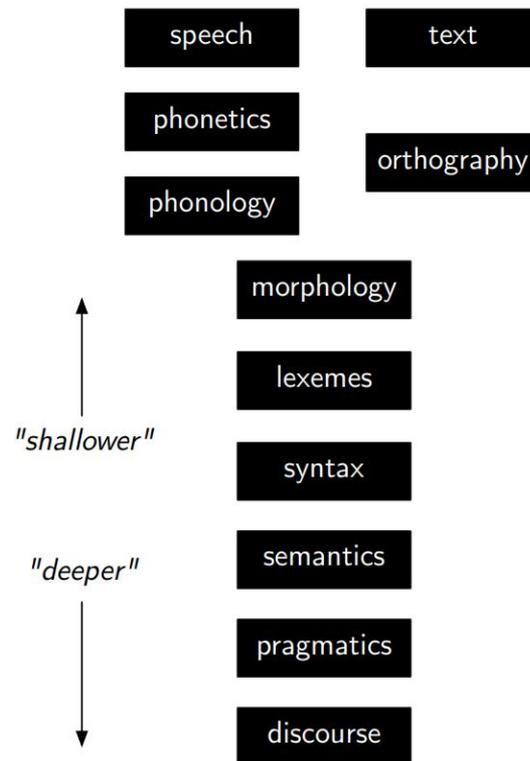
Tokens This is a simple sentence .

Morphology be
3sg
present

Syntax

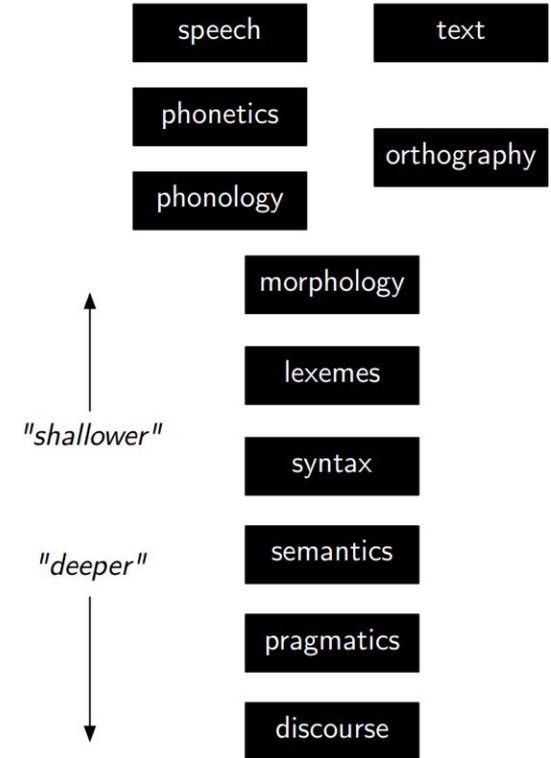
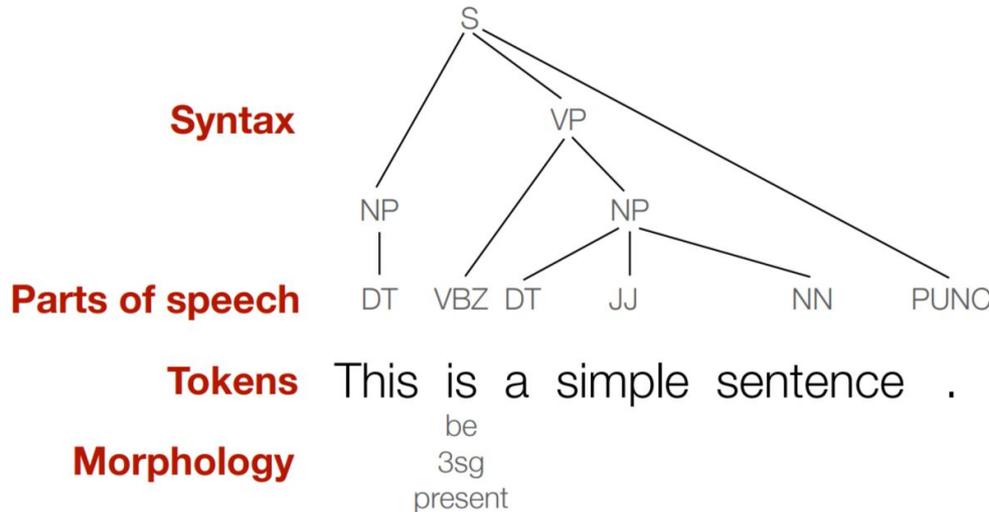
- Part-of-speech tagging

Parts of speech	DT	VBZ	DT	JJ	NN	PUNC
Tokens	This	is	a	simple	sentence	.
Morphology		be				
		3sg				
		present				



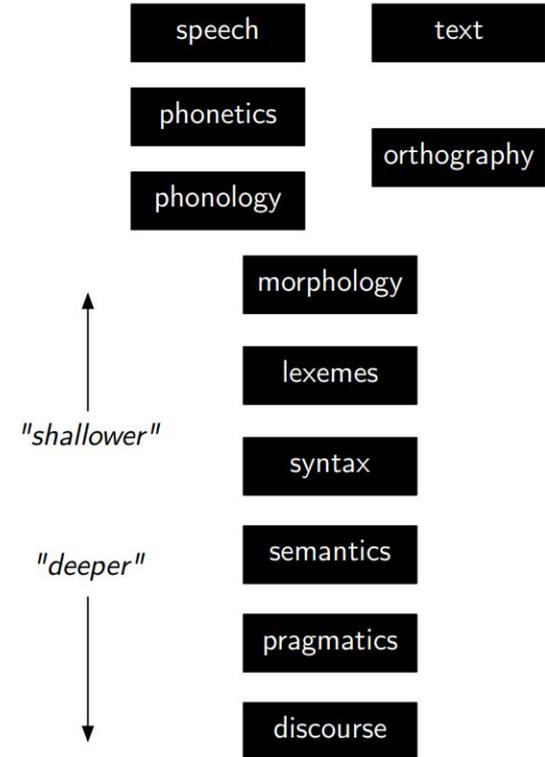
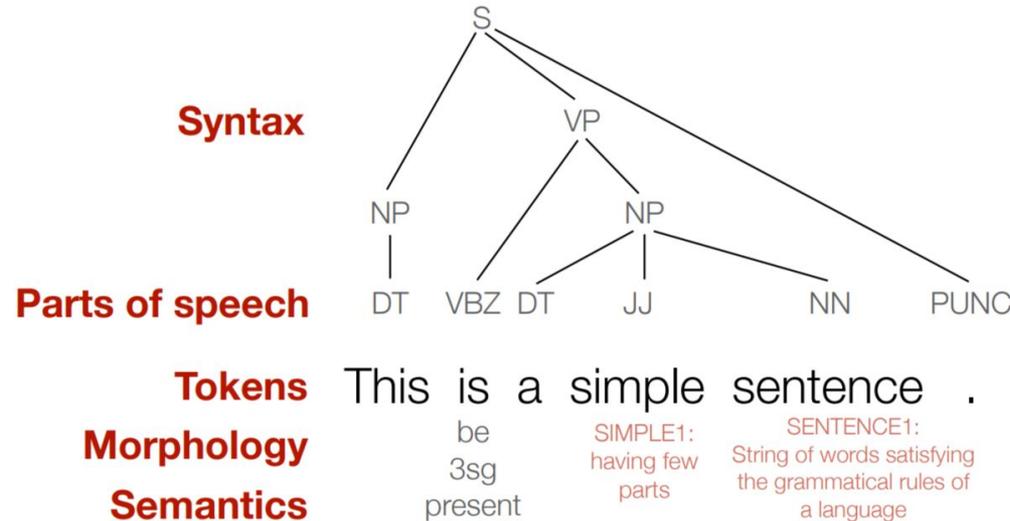
Syntax

- Part-of-speech tagging
- Syntactic parsing



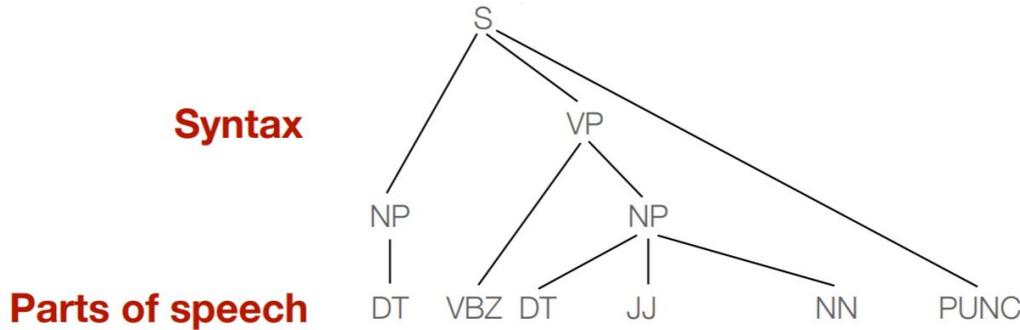
Semantics

- Named entity recognition
- Word sense disambiguation
- Semantic role labelling



Discourse

- Reference resolution
- Discourse parsing



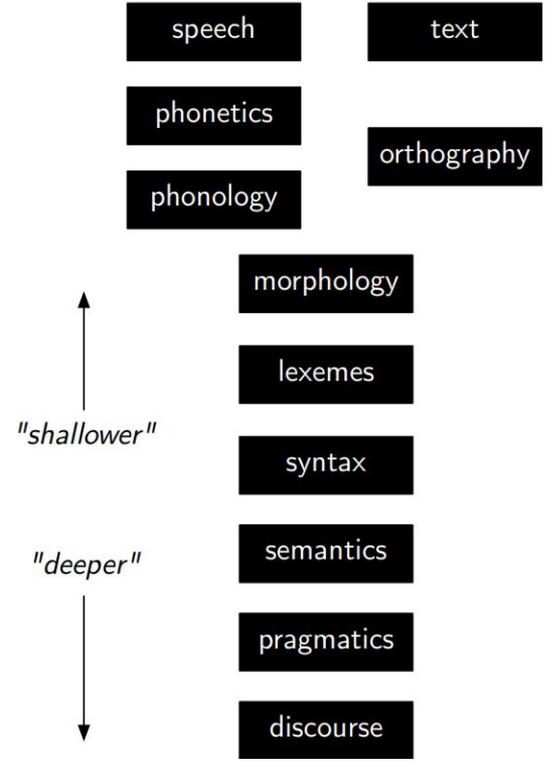
Tokens This is a simple sentence .

Morphology be SIMPLE1: having few parts SENTENCE1: String of words satisfying the grammatical rules of a language

Semantics 3sg present

Discourse But an instructive one .

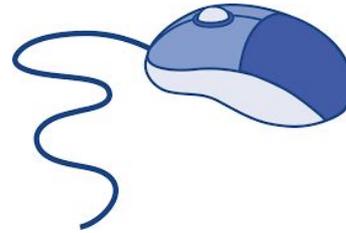
coreferent



Why is language interpretation hard?

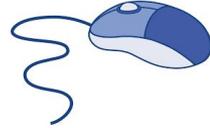
1. **Ambiguity**
2. Variation
3. Sparsity
4. Expressivity
5. Unmodeled variables
6. Unknown representation \mathcal{R}

Ambiguity: word sense disambiguation



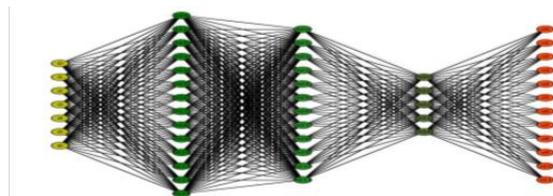
Ambiguity

- Ambiguity at multiple levels:
 - Word senses: **bank** (finance or river?)
 - Part of speech: **chair** (noun or verb?)
 - Syntactic structure: **I can see a man with a telescope**
 - Multiple: **I saw her duck**



Dealing with ambiguity

- How can we model ambiguity and choose the correct analysis in context?
 - non-probabilistic methods (FSMs for morphology, CKY parsers for syntax) return **all possible analyses**.
 - probabilistic models (HMMs for part-of-speech tagging, PCFGs for syntax) and algorithms (Viterbi, probabilistic CKY) return **the best possible analysis**, i.e., the most probable one according to the model
 - Neural networks, pretrained language models now provide end-to-end solutions



- But the “best” analysis is only good if our probabilities are accurate. Where do they come from?

Corpora

- A corpus is a collection of text
 - Often annotated in some way
 - Sometimes just lots of text
- Examples
 - Penn Treebank: 1M words of parsed WSJ
 - Canadian Hansards: 10M+ words of aligned French / English sentences
 - Yelp reviews
 - The Web: billions of words of who knows what

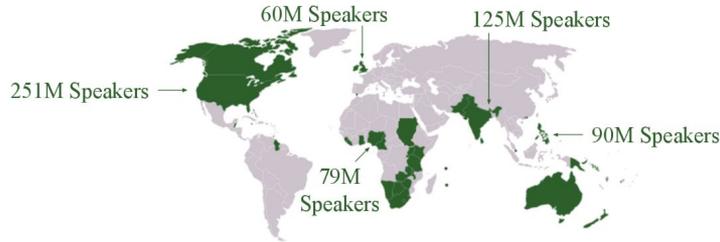


Why is language interpretation hard?

1. Ambiguity
2. **Variation**
3. Sparsity
4. Expressivity
5. Unmodeled variables
6. Unknown representation \mathcal{R}

Variation

- ~7K languages
- Thousands of language varieties



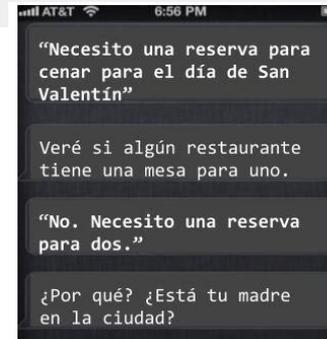
Englishes



Africa is a continent with a very high linguistic diversity: there are an estimated 1.5-2K African languages from 6 language families. **1.33 billion people**

NLP beyond English

- ~7,000 languages
- thousands of language varieties



Spanish
534 million speakers



Hindi
615 million speakers



Swahili
100 million speakers



American English



Scottish English

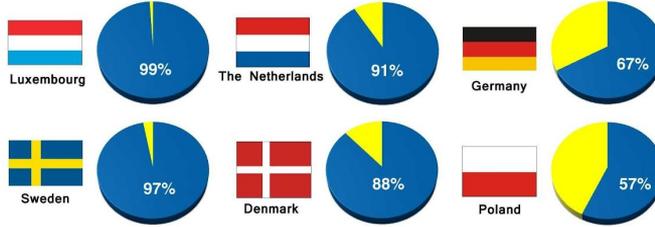


Hinglish

Most of the world today is multilingual

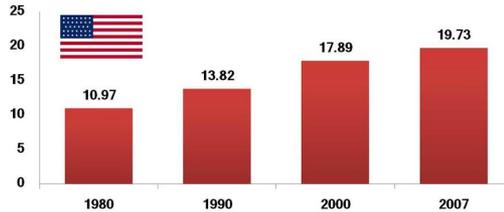
Percentage of Bilingual Speakers in the World

European Union



Source: European Commission, "Europeans and their Languages," 2006

Percentage of US Population who spoke a language other than English at home by year

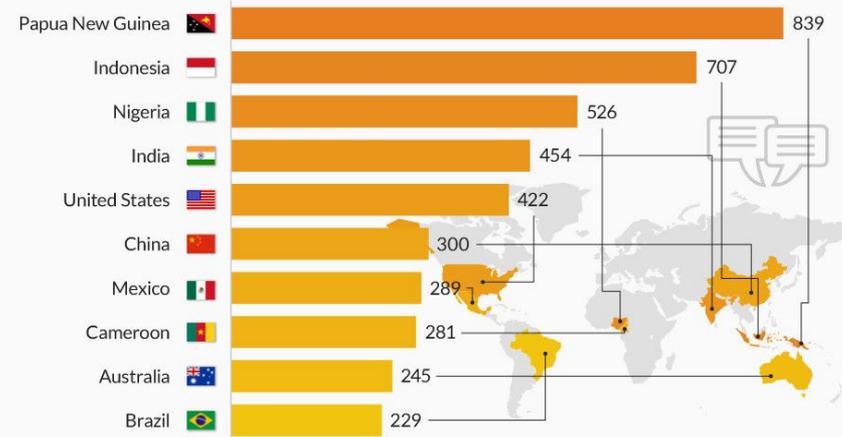


Source: U.S. Census Bureau, 2007 American Community Survey

Source: US Census Bureau

The Countries With The Most Spoken Languages

Number of living languages spoken per country in 2015



Source: Ethnologue

Tokenization

这是一个简单的句子

WORDS

This is a simple sentence

זה משפט פשוט

Tokenization + disambiguation

in tea
her daughter

בתה

- most of the vowels unspecified

in tea	בתה
in the tea	בהתה
that in tea	שבתה
that in the tea	שבהתה
and that in the tea	ושבהתה

ושבתה

and her saturday	ו+שבת+ה
and that in tea	ו+ש+ב+תה
and that her daughter	ו+ש+בת+ה

- most of the vowels unspecified
- particles, prepositions, the definite article, conjunctions attach to the words which follow them
- tokenization is highly ambiguous

Tokenization + morphological analysis

- Quechua

Much'ananyakapushasqakupuniñataqsunamá

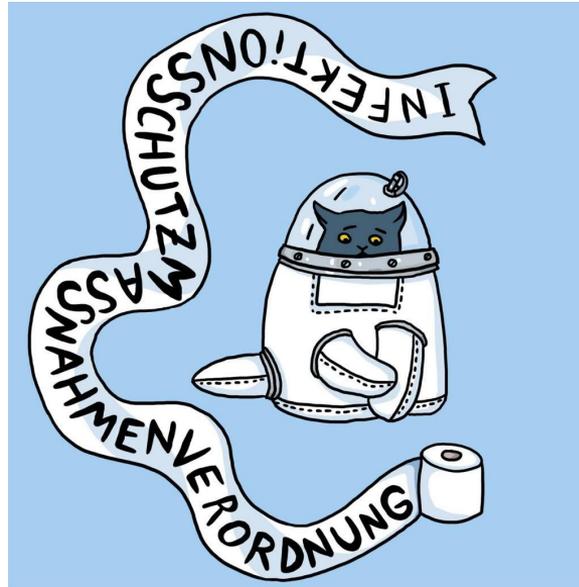
Much'a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

"So they really always have been kissing each other then"

Much'a	to kiss
-na	expresses obligation, lost in translation
-naya	expresses desire
-ka	diminutive
-pu	reflexive (kiss *eachother*)
-sha	progressive (kiss*ing*)
-sqa	declaring something the speaker has not personally witnessed
-ku	3rd person plural (they kiss)
-puni	definitive (really*)
-ña	always
-taq	statement of contrast (...then)
-suna	expressing uncertainty (So...)
-má	expressing that the speaker is surprised

Tokenization + morphological analysis

- German



Infektionsschutzmaßnahmenverordnung

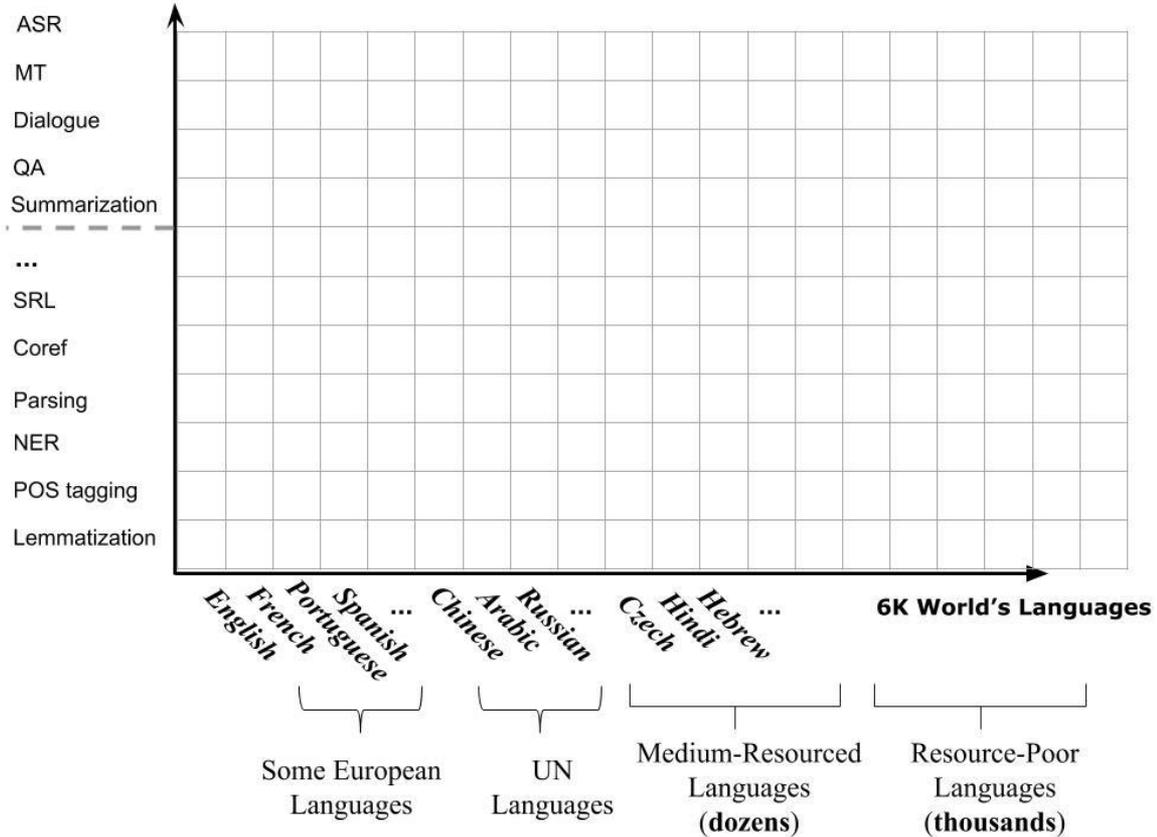
Semantic analysis

- Every language sees the world in a different way
 - For example, it could depend on cultural or historical conditions



- Russian has very few words for colors, Japanese has hundreds
- Multiword expressions, e.g. [happy as a clam](#), [it's raining cats and dogs](#) or [wake up](#) and metaphors, e.g. [love is a journey](#) are very different across languages

NLP Technologies/Applications



Linguistic variation

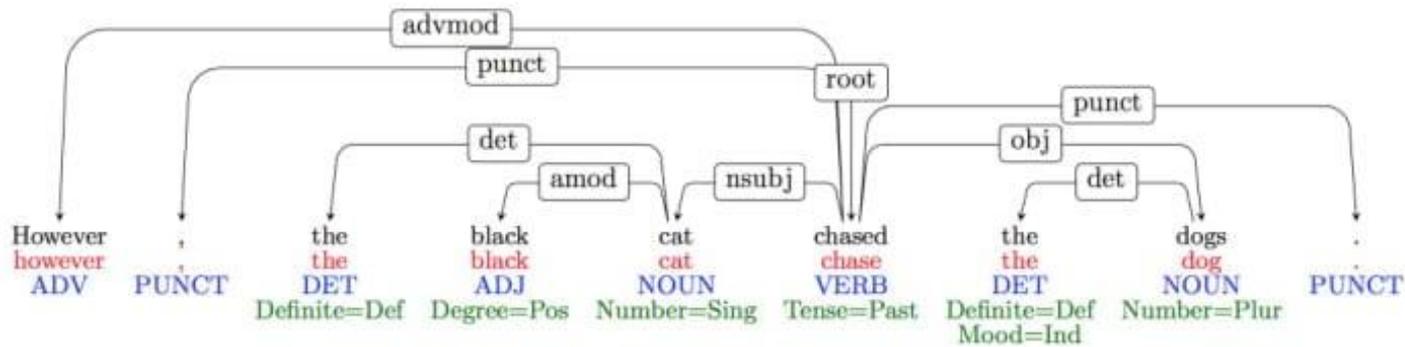
- Non-standard language, emojis, hashtags, names



chowdownwithchan #crab and #pork #xiaolongbao at @dintaifungusa... where else? 🤔👩 Note the cute little crab indicator in the 2nd pic 🦀💕

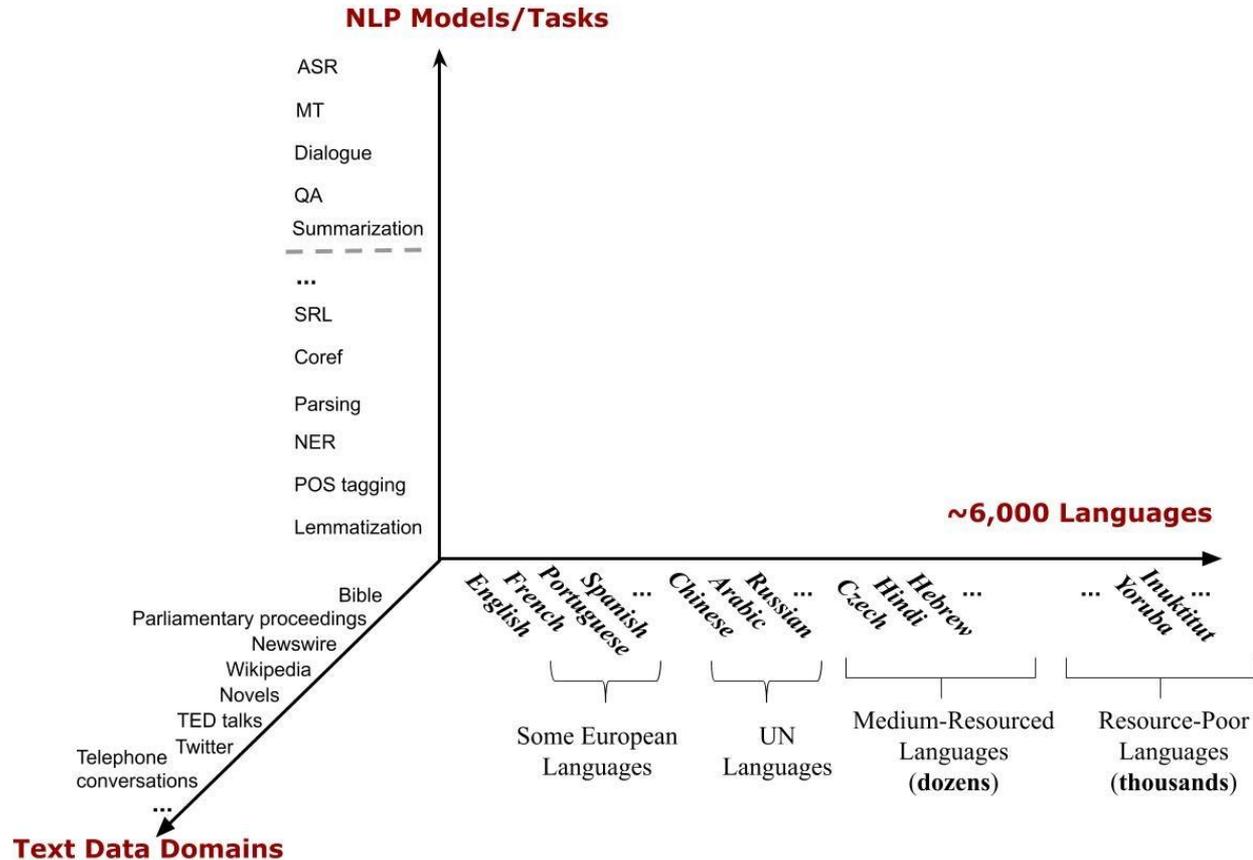
Variation

- Suppose we train a part of speech tagger or a parser on the Wall Street Journal



- What will happen if we try to use this tagger/parser for social media??

@_rkpntrnte hindi ko alam babe eh, absent ako
kanina I'm sick rn hahaha 🤔🙌



Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation \mathcal{R}

Sparsity

Sparse data due to **Zipf's Law**

- To illustrate, let's look at the frequencies of different words in a large text corpus
- Assume “word” is a string of letters separated by spaces

Word Counts

Most frequent words in the English Europarl corpus (out of 24m word tokens)

any word		nouns	
Frequency	Token	Frequency	Token
1,698,599	the	124,598	European
849,256	of	104,325	Mr
793,731	to	92,195	Commission
640,257	and	66,781	President
508,560	in	62,867	Parliament
407,638	that	57,804	Union
400,467	is	53,683	report
394,778	a	53,547	Council
263,040	I	45,842	States

Word Counts

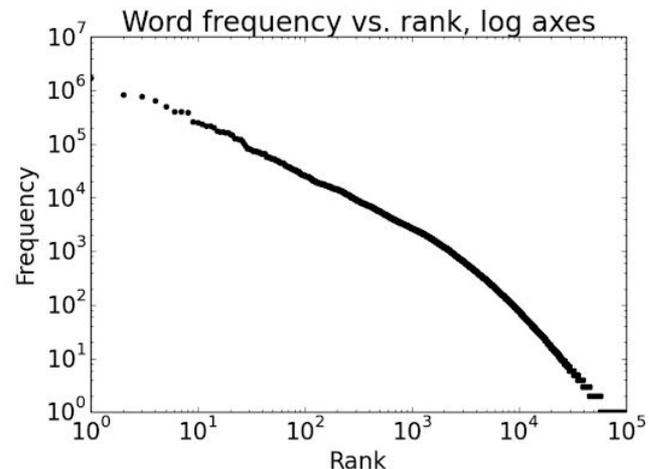
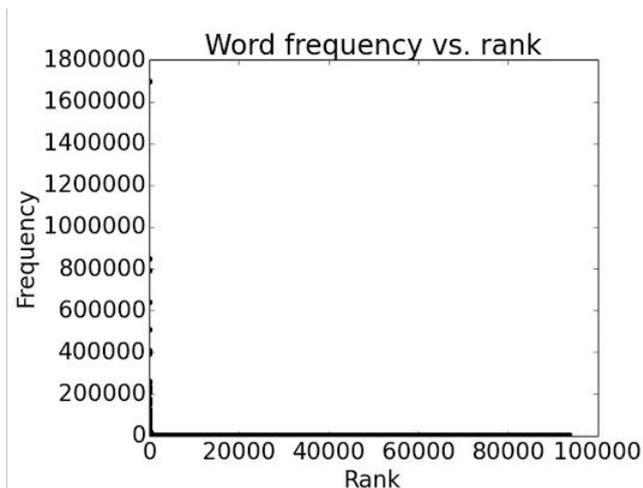
But also, out of 93,638 distinct words (word types), 36,231 occur only once.

Examples:

- cornflakes, mathematicians, fuzziness, jumbling
- pseudo-rapporteur, lobby-ridden, perfunctorily,
- Lycketoft, UNCITRAL, H-0695
- policyfor, Commissioneris, 145.95, 27a

Plotting word frequencies

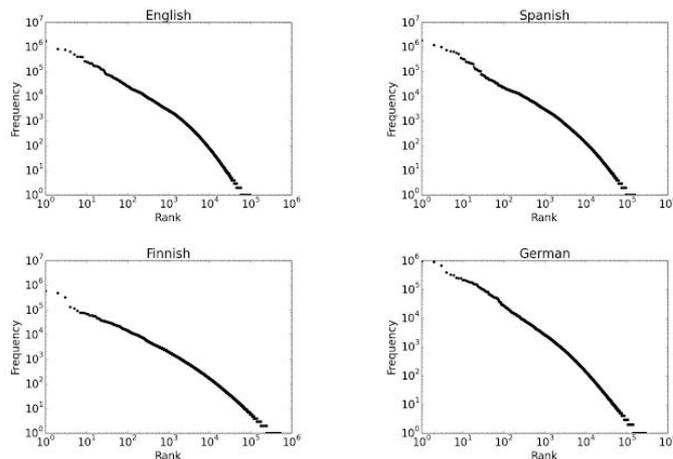
Order words by frequency. What is the frequency of n th ranked word?



Zipf's Law

Implications

- Regardless of how large our corpus is, there will be a lot of infrequent (and zero-frequency!) words
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen



Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation \mathcal{R}

Expressivity

Not only can one form have different meanings (ambiguity) but the same meaning can be expressed with different forms:

She gave the book to Tom vs. She gave Tom the book

Some kids popped by vs. A few children visited

Is that window still open? vs. Please close the window

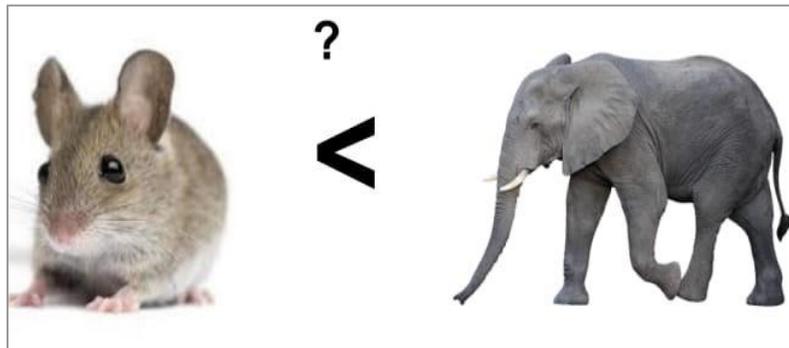
Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. **Unmodeled variables**
7. Unknown representation \mathcal{R}

Unmodeled variables



“Drink this milk”



World knowledge

- I dropped the glass on the floor and it broke
- I dropped the hammer on the glass and it broke

Why is language interpretation hard?

1. Ambiguity
2. Scale
3. Variation
4. Sparsity
5. Expressivity
6. Unmodeled variables
7. Unknown representation \mathcal{R}

Unknown representation

- Very difficult to capture *what is \mathcal{R}* , since we don't even know how to represent the knowledge a human has/needs:
 - What is the “meaning” of a word or sentence?
 - How to model context?
 - Other general knowledge?

Desiderata for NLP models

- **Sensitivity** to a wide range of phenomena and constraints in human language
- **Generality** across languages, modalities, genres, styles
- Strong **formal guarantees** (e.g., convergence, statistical efficiency, consistency)
- **High accuracy** when judged against expert annotations or test data
- **Ethical**

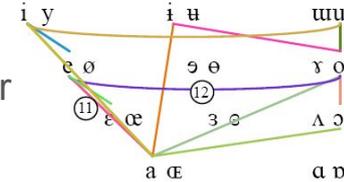
NLP $\stackrel{?}{=}$ Machine Learning

- To be successful, a machine learner needs bias/assumptions; for NLP, that might be linguistic theory/representations.
- Symbolic, probabilistic, and connectionist ML have all seen NLP as a source of inspiring applications.

What is nearby NLP?

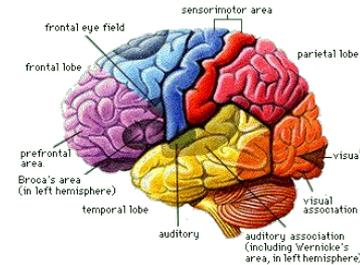
- Computational Linguistics

- Using computational methods to learn more about how language works
- We end up doing this and using it



- Cognitive Science

- Figuring out how the human brain works
- Includes the bits that do language
- Humans: the only working NLP prototype!



- Speech Processing

- Mapping audio signals to text
- Traditionally separate from NLP, converging?
- Two components: acoustic models and language models
- Language models in the domain of stat NLP



Text Classification

Is this spam?

from: **ECRES 2022 <2022@ecres.net>** [via](#) amazonses.com
reply-to: 2022@ecres.net
to: yuliats@cs.washington.edu
date: Feb 22, 2022, 7:21 AM
subject: The Best Renewable Energy Conference (Last chance !)
signed-by: amazonses.com
security: Standard encryption (TLS) [Learn more](#)

Dear Colleague,

Account: yuliats@cs.washington.edu

Good news: Due to many requests, the submission deadline has been extended to **10 March 2022** (It is firm date).

We would like to invite you to submit a paper to 10. European Conference on Renewable Energy Systems (ECRES). **ECRES 2022 will be held hybrid mode, the participants can present their papers physically or online.** The event is going to be organized in Istanbul/Turkey under the technical sponsorship of Istanbul Medeniyet University and many international institutions. The conference is highly international with the participants from all continents and more than 40 countries.

The submission deadline and special and regular issue journals can be seen in [ecres.net](#)

There will be keynote speakers who will address specific topics of energy as you would see at [ecres.net/keynotes.html](#)

[CLICK FOR PAPER SUBMISSION](#)

All accepted papers will be published in a special Conference Proceedings under a specific ISBN. Besides, the extended versions will be delivered to reputable journals **indexed in SCI, E-SCI, SCOPUS, and EBSCO**. You can check our previous journal publications from [ecres.net](#) . **Please note that the official journal of the event, [Journal of Energy Systems \(\[dergipark.org.tr/jes\]\(http://dergipark.org.tr/jes\) \) is also indexed in SCOPUS.](#)**

Spam classification

Dear Colleague,

Account: yuliats@cs.washington.edu

Good news: Due to many requests, the submission deadline has been extended to 10 March 2022 (It is firm date).

We would like to invite you to submit a paper to the conference on Renewable Energy Systems (ECRES). **ECRES 2022** will be held in Istanbul/Turkey under the technical sponsorship of Medeniyet University and many international institutions. The conference is international with the participants from all continents and more than 40 countries.



The submission deadline and special and regular issue journals can be seen in ecres.net

There will be keynote speakers who will address specific topics of energy as you would see at ecres.net/keynotes.html

[CLICK FOR PAPER SUBMISSION](#)

All accepted papers will be published in a special Conference Proceedings under a specific ISBN. Besides, the extended versions will be delivered to reputable journals indexed in **SCI, E-SCI, SCOPUS, and EBSCO**. You can check our previous journal publications from ecres.net. **Please note that the official journal of the event, Journal of Energy Systems (dergipark.org.tr/jes) is also indexed in SCOPUS.**

Invitation to present at the February 2022 Wikimedia Research Showcase



Emily Lescak <elescak@wikimedia.org>
to yuliats@cs.washington.edu

Hi Yulia,

I am a member of the [Research Team](#) at the Wikimedia Foundation. On behalf of the team, I would like to invite you to present your research on social biases on Wikipedia at our [Research Showcase](#) in February 2022. This topic fits into our theme for this showcase, which is gaps and biases on Wikipedia.

The Wikimedia Research Showcase is a monthly, public lecture series where Foundation, academic, and community members present their work related to Wikipedia, Wikimedia, peer production, and open-source software. We focus on topics and projects that we think our audience—a global community of academic researchers, Wikipedia editors, and Wikimedia staff—would find interesting and relevant to their work.

Research Showcase presentations are generally 20 minutes long, with an additional 10 minutes for questions. We invite two presenters to every showcase. Most presenters choose to use slides to present their work.

The February showcase takes place on the 16th at 9:15AM Pacific / 17:15 UTC. You can watch past showcases on our [YouTube](#) and also archived for later viewing on the [Wikimedia Foundation's YouTube channel](#)

If this date does not work for you, but you are still interested in giving a showcase, please let us know so we can discuss other options.

I hope to get a chance to see your work presented at the Research Showcase!

Sincerely,

Emily

--



Language ID

Аяны замд түр зогсон тэнгэрийн байдлыг ажиглаад хөдлөх зуур гутал дор шинэхэн орсон цас шаржигнан дуугарч байв. Цасны тухай бодол сонин юм. Хот хүрээ тийш цас орвол орно л биз гэсэн хэнэггүй бодол маань хөдөө талд, говийн ээрэм хөндийд, малын бэлчээрт, малчдын хотонд болохоор солигдож эргэцүүлэн бодох нь хачин. Цас хэр орсон бол?

Београд, 16. јун 2013. године – Председник Владе Републике Србије Ивица Дачић честитао је кајакашици златне медаље у олимпијској дисциплини К-1, 500 метара, као и у двоструко дужој стази освојене на првенству Европе у Португалији.

Beograd, 16. јun 2013. године – Председник Владе Републике Србије Ivica Dačić čestitao је кајакашци златне медаље у олимпијској дисциплини К-1, 500 метара, као и у двојструко дузој стази освојене на првенству Европе у Португалији.

Nestranski Urad za vladno odgovornost ZDA je objavil eksplozivno mnenje, da je vlada predsednika Donalda Trumpa kršila zvezno zakonodajo, ko je zadrževala izplačilo kongresno potrjene vojaške pomoči Ukrajini zaradi političnih razlogov. Predstavniški dom kongresa je prav zaradi tega sprožil ustavno obtožbo proti Trumpu.

Language ID

Аяны замд түр зогсон тэнгэрийн байдлыг ажиглаад хөдлөх зуур гутал дор шинэхэн орсон цас шаржигнан дуугарч байв. Цасны тухай бодол сонин юм. Хот **mongolian** рвол орно л биз гэсэн хэнэггүй бодол маань хөдөө тал **mongolian** өндийд, малын бэлчээрт, малчдын хотонд болохоор солигдож эргэцүүлэн бодох нь хачин. Цас хэр орсон бол?

Београд, 16. јун 2013. године – Председник Владе Републике Срб **serbian** неститао је кајакашици златне медаље у оли **serbian** ини K-1, 500 метара, као и у двоструко дужој стази освојене на првенству Европе у Португалији.

Beograd, 16. јun 2013. године – Председник Владе Републике Ср **serbian** итао је кајакашици златне медаље у о **serbian** K-1, 500 метара, као и у двојструко дуђој стази освојене на првенству Европе у Португалији.

Nestranski Urad za vladno odgovornost ZDA je objavil eksplozivno mnenje, da je vlada predsednika Donalda Trumpa kršila zvezno zakonodajo, ko je zadrževala izplačilo k **slovenian** vojaške pomoči Ukrajini zaradi političnih razlogov. Predstavniški d **slovenian** av zaradi tega sprožil ustavno obtožbo proti Trumpu.

Sentiment analysis



By [John Neal](#)

This review is from: [Accoutrements Horse Head Mask \(Toy\)](#)

When I turned State's Witness, they didn't have enough money to put me in the Witness Protection Program, so they bought me this mask and gave me a list of suggested places to move. Since then I've lived my life in peace and safety knowing that my old identity is forever obscured by this life-saving item.



By [Christine E. Torok](#)

Verified Purchase ([What's this?](#))

First of all, for taste I would rate these a 5. So good. Soft, true-to-taste fruit flavors like the sugar variety...I was a happy camper.

BUT (or should I say BUTT), not long after eating about 20 of these all hell broke loose. I had a gastrointestinal experience like nothing I've ever imagined. Cramps, sweating, bloating beyond my worst nightmare. I've had food poisoning from some bad shellfish and that was almost like a skip in the park compared to what was going on inside

Sentiment analysis



By [John Neal](#)

This review is from: [Accoutrements Horse Head Mask \(Toy\)](#)

When I turned State's Witness, they didn't have enough money to put me in the Witness Protection Program, so they bought me this mask and I came along for sugar and places to move. Since then, I've lived my life in peace and safety knowing that my old identity is forever obscured by this life-saving item.



By [Christine E. Torok](#)

Verified Purchase ([What's this?](#))

First of all, for taste I would rate these a 5. So good. Soft, true-to-taste fruit flavors like the sugar variety...I was a happy camper.

BUT (or should I say BUTT), not long after eating about 20 of these all hell broke loose. I had a gastrointestinal experience like nothing I've ever imagined. Cramps, sweating, floating beyond my worst nightmare. I've had food poisoning from some bad shellfish and that was almost like a skip in the park compared to what was going on inside



