AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 771: Computer Processing of Natural Language
Lecture 2: Languages: Text, Sound, Visual, Mixed

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

23RD AUG 2022

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 2

- Introduction Section
  - Opening comments
  - Kite tool
- Main Section
- Concluding Section
  - Course Project
  - About Next Lecture – Lecture 3

Main Section

- What is a language?
  - Media: Text, Sound (speech), Visual (image, video), Multi (modal, media)
  - Media representation
- Processing data
  - Reading
  - Searching content fragment, Manipulating content
  - Writing
- Ethical considerations
- Concluding comments

Opening Comments

- Sharing information with instructor
  - Google sheet shared on Piazza
  - Sharing code
    - Create private Github repo named **csce771-fall2022-<yourname>**
    - Share with instructor only (biplav-s)
  - Sharing docs (quizzes and report):
    - Create Google drive named **CSCE Fall2022-<YourName>**
    - Share with instructor only (firstname.lastname@gmail.com)
    - Sub-folders for quizzes and paper-report
      - Sub-folder like Quiz1 under quizzes
- About code resources
  - Github site of instructor's code fragments: https://github.com/biplav-s/course-nl-f22/
  - Other material
    - Introduction to NLP Webinar by WomenWhoCode group: https://github.com/WomenWhoCode/WWCodeDataScience/tree/master/Intro_to_NLP
    - CSCE 771, Fall 2020 - https://github.com/biplav-s/course-nl/
-

Demonstration: Unsupervised Text Exploration

- **Input**: a document, i.e., a piece of text or URL
- **Output**: what information does the document convey ?

- **Context**: no assumption about what the document contains, but have common sense assumptions about the domains in the world like people, politics, science.

- **Tool**: Kite
  http://casy.cse.sc.edu/kite

http://casy.cse.sc.edu/kite

1

Distinguished Professor Emeritus

Michael N. Huhns

NCR Chair Emeritus of Computer Science and Engineering

Detailed Resume | Research Interests | Publications

Dr. Michael N. Huhns is the NCR Distinguished Professor Emeritus of Computer Science and Engineering at the University of South Carolina. Prior to this, he was the Chair of the Department of Computer Science and Engineering. He is a Fellow of the IEEE, a Fellow of the Association for the Advancement of Artificial Intelligence, and a Senior Member of the ACM.

CONTACT INFORMATION

Department of Computer Science and Engineering
Swearingen Engineering Center
University of South Carolina
Columbia, SC 29208
huhns@sc.edu
Phone: +1 (803) 777-5921
Fax: +1 (803) 777-3767

BRIEF BIOGRAPHY (a detailed resume is here)

Dr. Huhns received the B.S.E.E. degree in 1969 from the University of Michigan, Ann Arbor, and the M.S. and Ph.D. degrees in electrical engineering in 1971 and 1975, respectively, from the University of Southern California, Los Angeles.

Before becoming a professor of computer science and engineering at the University of South Carolina, he conducted research on the Argo, Antares, RAD, Carnot, and InfoSleuth projects at the Microelectronics and Computer Technology Corporation as a Senior Member of the Research Division. He was also an adjunct professor in computer sciences at the University of Texas. Prior to joining MCC, he was an associate professor at the University of South Carolina, a research assistant in image processing at the University of Southern California, and a radar systems engineer at Hughes Aircraft Company.

2

AN UNSUPERVISED SYSTEM FOR TEXT EXPLORATION - KITE

Enter link...

SUBMIT

OR, Try Some Suggested URLs

- Computer : https://www.geeksforgeeks.org/compiler-construction-tools/
- Politics : http://lite.cnn.com/en/article/h_f68f79dede493cb0a82201f6ea0ce293
- Science : https://vikaspedia.in/education/childrens-corner/science-section/articles-on-science
- Recreation : https://www.nytimes.com/2022/05/23/sports/baseball/roger-angell.html
- RFP : https://beta.nsf.gov/funding/opportunities/division-chemistry-disciplinary-research-programs-no-deadline-pilot-che
- Water Regulations : https://www.epa.gov/dwreginfo/public-water-system-supervision-program-water-supply-guidance-manual
- Person : https://www.cse.sc.edu/~huhns/

SUBMIT

3

AN UNSUPERVISED SYSTEM FOR TEXT EXPLORATION - KITE

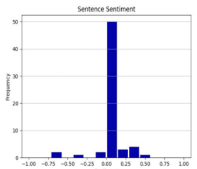
Overview | Top Entities | Top Events | Researcher | Full Text

Overview Toggle

Statistics

Number of words: 712

Number of sentences: 63



Domain: Person (Michael N. Huhns)

4

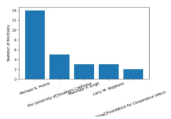
Top Sentences Toggle

Positive Sentences | Negative Sentences

Sentence	Sentiment
Before becoming a professor of computer science and engineering at the University of South Carolina, he conducted research on the Argo,	0.45
Detailed Resume | Research Interests | Publications	0.4
Dr. Huhns is an associate editor for the Journal of Autonomous Agents and Multi-Agent Systems , and IEEE Internet Computing .	0.4
He is on the editorial boards of the International Journal of Cooperative Information Systems, Journal of Intelligent Manufacturing, and Journal of Emerging Mechanical Engineering Technology .	0.4
Robustness and Social Good via Widespread Multiagent Development ," in Proc.	0.37

Top Entities Toggle

Summary | Michael N. Huhns | the University of Southern California | Munindar P. Singh | Larry M. Stephens | the International Foundation for Cooperative Information Systems



Michael N. Huhns

the University of Southern California

Munindar P. Singh

Larry M. Stephens

the International Foundation for Cooperative Information Systems

Top Events Toggle

Date	Location	Verb	Sentence
1969	-	received	Dr. Huhns received the B.S.E.E. degree in 1969 from the University of Michigan , Ann Arbor , and the M.S. and Ph.D. degrees in electrical engineering in 1971 and 1975 , respectively , from the University of Southern California , Los Angeles .
-	-	Distributed	He is the author of over 250 technical papers in machine intelligence and an editor of the books Distributed Artificial Intelligence , Volumes
-	-	distributed	My research has ranged broadly over many areas of information technology , including service - oriented computing , distributed artificial intelligence and multiagent systems , machine learning , computer vision , enterprise integration , and computational social systems .
-	-	distributing	Its focus , however , has been on the three complementary themes of (1) distributing the loci of computational intelligence , (2) enhancing the intelligence of the distributed computations , and (3) increasing the effectiveness of

5

Full Text Toggle

Distinguished Professor Emeritus Michael N. Huhns PERSON NCR Chair ORG Emeritus of Computer Science and Engineering ORG Detailed Resume | Research Interests | Publications Dr. Michael N. Huhns PERSON is the NCR ORG Distinguished Professor Emeritus of Computer Science and Engineering ORG at the University of South Carolina ORG . Prior to this, he was the Chair of the Department of Computer Science and Engineering ORG . He is a Fellow of the IEEE ORG , a Fellow of the Association for the Advancement of Artificial Intelligence ORG , and a Senior Member of the ACM ORG . CONTACT INFORMATION Department of Computer Science and Engineering Swearingen Engineering Center University of South Carolina Columbia ORG , SC ORG 29208 CARDINAL huhns@sc.edu Phone: +1 (803 CARDINAL) 777-5921 Fax: +1 (803 CARDINAL) 777-3767 PRODUCT BRIEF BIOGRAPHY (a detailed resume is here) Dr. Huhns PERSON received the B.S.E.E. GPE degree in 1969 DATE from the University of Michigan ORG , Ann Arbor GPE , and the M.S.

6

Researcher Toggle

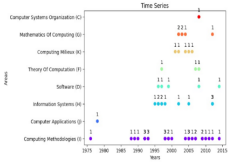
Name: Michael N. Huhns

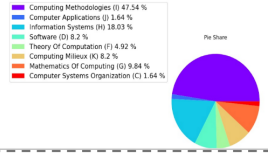
ACM | JEL

Summary of top ACM Classification Codes

Code	Term	Mentions
I.2.11	Distributed Artificial Intelligence	5
H	Information Systems	12
J.2	Physical Sciences And Engineering	8
G.4	Mathematical Software	6
I.5.4	Applications	3
F.1.1	Models Of Computation	3
H.2.5	Heterogeneous Databases	3
I.2.3	Deduction And Theorem Proving	2

Areas of Interest





7

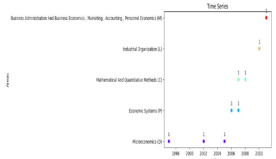
Name: Michael N. Huhns

ACM | JEL

Summary of top JEL Classification Codes

Code	Term	Mentions
O18	Urban, Rural, Regional, And Transportation Analysis ; Housing ; Infrastructure	2
L8	Industry Studies; Services	2
C52	Model Evaluation, Validation, And Selection	2
P11	Planning, Coordination, And Reform	2
L14	Transactional Relationships ; Contracts And Reputation ; Networks	1
M14	Corporate Culture ; Diversity ; Social Responsibility	1
D83	Search ; Learning ; Information And Knowledge ; Communication ; Belief ; Unawareness	1
D04	Microeconomic Policy, Formulation, Implementation, And Evaluation	1
D3	Distribution	1

Areas of Interest



8

Recent Publications

Year	Title
2021	Sociotechnical Perspectives on AI Ethics and Accountability
2021	Theme Articles
2020	Superconcept formation system—An ontology matching algorithm for service discovery
2018	IEEE Internet Computing Published by the IEEE Computer Society 1089-7801/18/$33.00© 2018 IEEE
2017	A service computing manifesto: the next 10 years

https://indianexpress.com/section/sports/cricket/

https://indianexpress.com/section/sports/cricket/

Home / Sports / Cricket

SPORTS CWG2022 CRICKET FOOTBALL TENNIS WWE PHOTOS PODCAST EXPLAINED

1

Cricket

It's tough not being India regular but I prefer staying positive: Sanju Samson

AUGUST 22, 2022 6:54:10 PM

Having shot to prominence as a teenager and then making International debut in a T20I against Zimbabwe here in 2015, the Kerala cricketer has featured only in seven ODIs and 16 T20Is till date.

Losing Shaheen Afridi before Asia Cup is a big setback for Pakistan: Saqlain

AUGUST 22, 2022 6:51:03 PM

Saqlain, a former Test off-spinner, said overall Pakistan had a compact bowling attack capable of doing well against all teams.

England tour of Pakistan: Rawalpindi, Multan, Karachi to host 3 Tests

AUGUST 22, 2022 3:00:40 PM

This will be England's second half of the tour, following seven T20Is in Karachi and Lahore from September 20 to October 2.

casy.cse.sc.edu/display/ 120%

Overview Top Entities Top Events Full Text

2

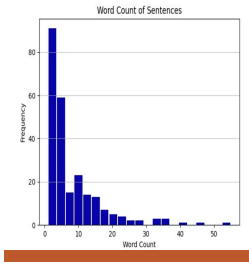
Overview Toggle

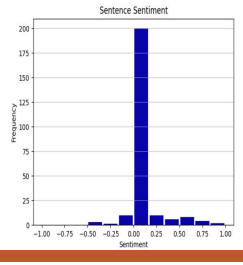
Statistics

Number of words: 1756

Number of sentences: 244

Domain: Recreation





casy.cse.sc.edu/display/ 120%

Top Sentences Toggle

Positive Sentences Negative Sentences

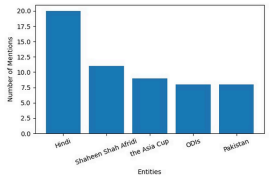
3

Sentence	Sentiment
6:48:47 pm India's 2022 Asia Cup opener against Pakistan will be a 14th meeting between the two in a competition that has seen some of their best.	1.0
Sportspersons greet the nation on Independence Day Advertisement Best of Express Aruna,	1.0
I am really proud to be a Malayali and represent my country as a Malayali cricketer, says Sar Samson August 22, 2022	
helped Pakistan beat the Netherlands by nine runs to win the ODI series.	
win Sports	

4

casy.cse.sc.edu/display/ 80%

Summary Hindi Shaheen Shah Afridi the Asia Cup ODIs Pakistan



Hindi

Shaheen Shah Afridi

the Asia Cup

ODIs

Pakistan

Top Events Toggle

Date	Location	Verb	Sentence
-	-	ruled	Shaheen Shah Afridi has been ruled out of the Asia Cup due to knee injury .
August 22, 2022	Sri Lanka	ruled	Mohammad Hasnain to replace injured Shaheen Shah Afridi in Pakistan squad August 22 , 2022 1:29:55 pm Shaheen Shah Afridi has been ruled out from due to the knee injury he sustained during the first Test match against Sri Lanka in Galle in July .
-	-	replace	Mohammad Hasnain to replace injured Shaheen Shah Afridi in Pakistan squad August 22 , 2022 1:29:55 pm Shaheen Shah Afridi has been ruled out from due to the knee injury he sustained during the first Test match against Sri Lanka in Galle in July .
August 22, 2022	India	reduce	August 22 , 2022 6:56:37 pm Injury - forced absence of Bumrah and Afridi will reduce venom of bowling attacks of India , Pakistan at the Asia Cup .
-	-	beat	helped Pakistan beat the Netherlands by nine runs to win the ODI series .

Blue: Subject, Green: Direct Object, Red: Verb

Full Text Toggle

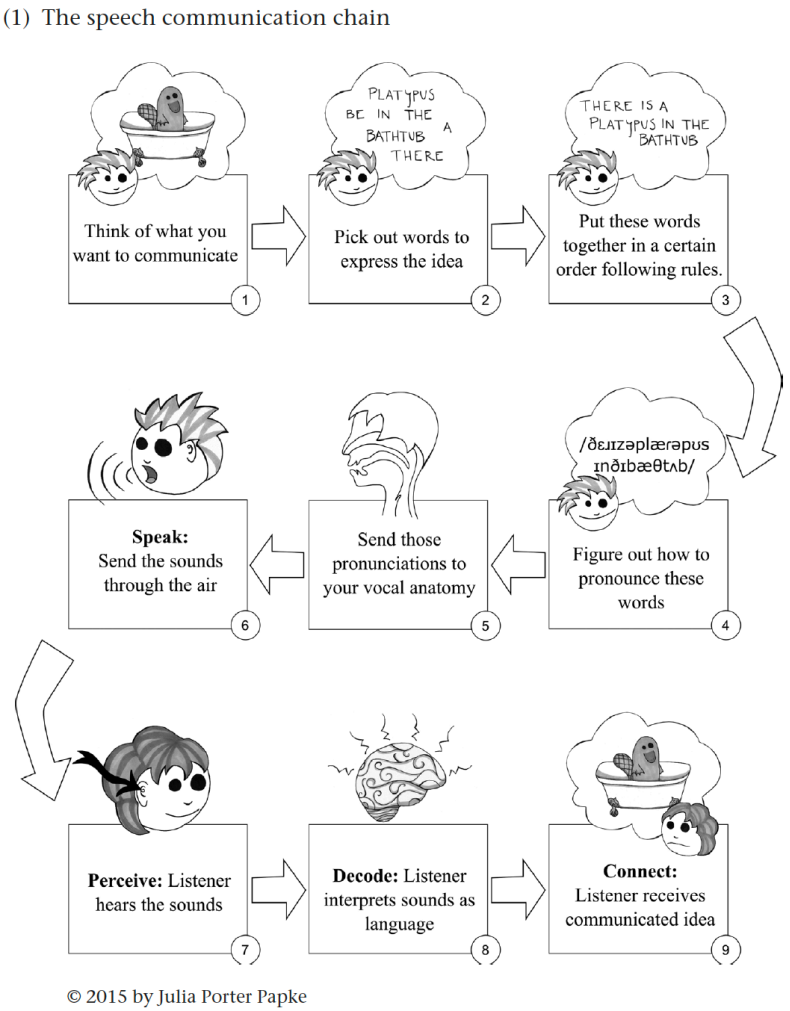
Exciting References

- Data statistics and preliminaries
  - https://movableink.com/blog/29-incredible-stats-that-prove-the-power-of-visual-marketing
  - https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=2c1ec9d60ba9
  - https://www.simplilearn.com/what-is-data-processing-article

- Artificial intelligence system learns concepts shared across video, audio, and text,
  - https://news.mit.edu/2022/ai-video-audio-text-connections-0504

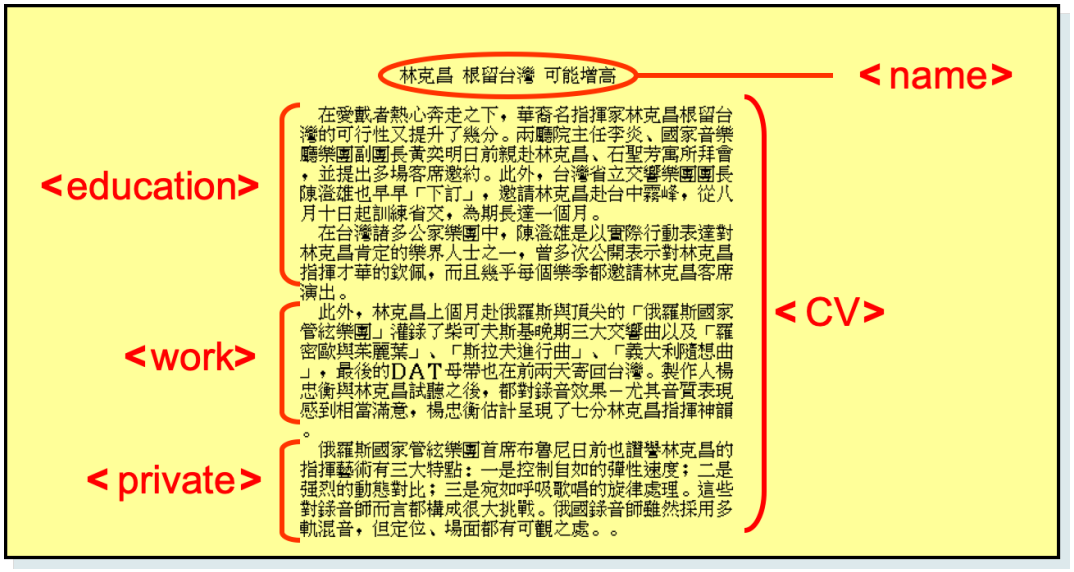
What is a Language?

Communication and Language



Case of a Computer Language

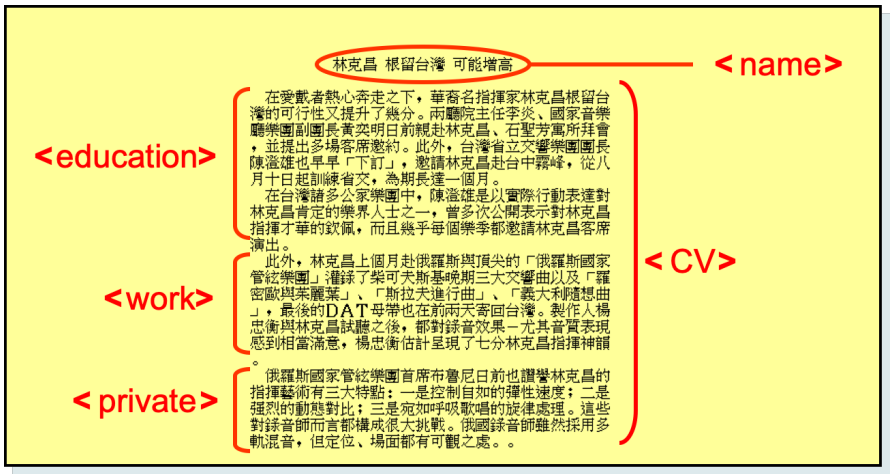
Example: XML

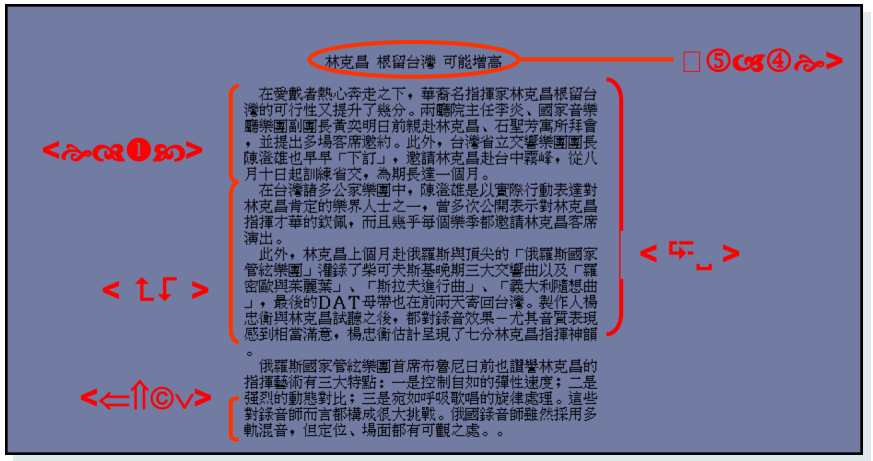


Slide Courtesy: Jim Hendler

Inter-Computer Communication

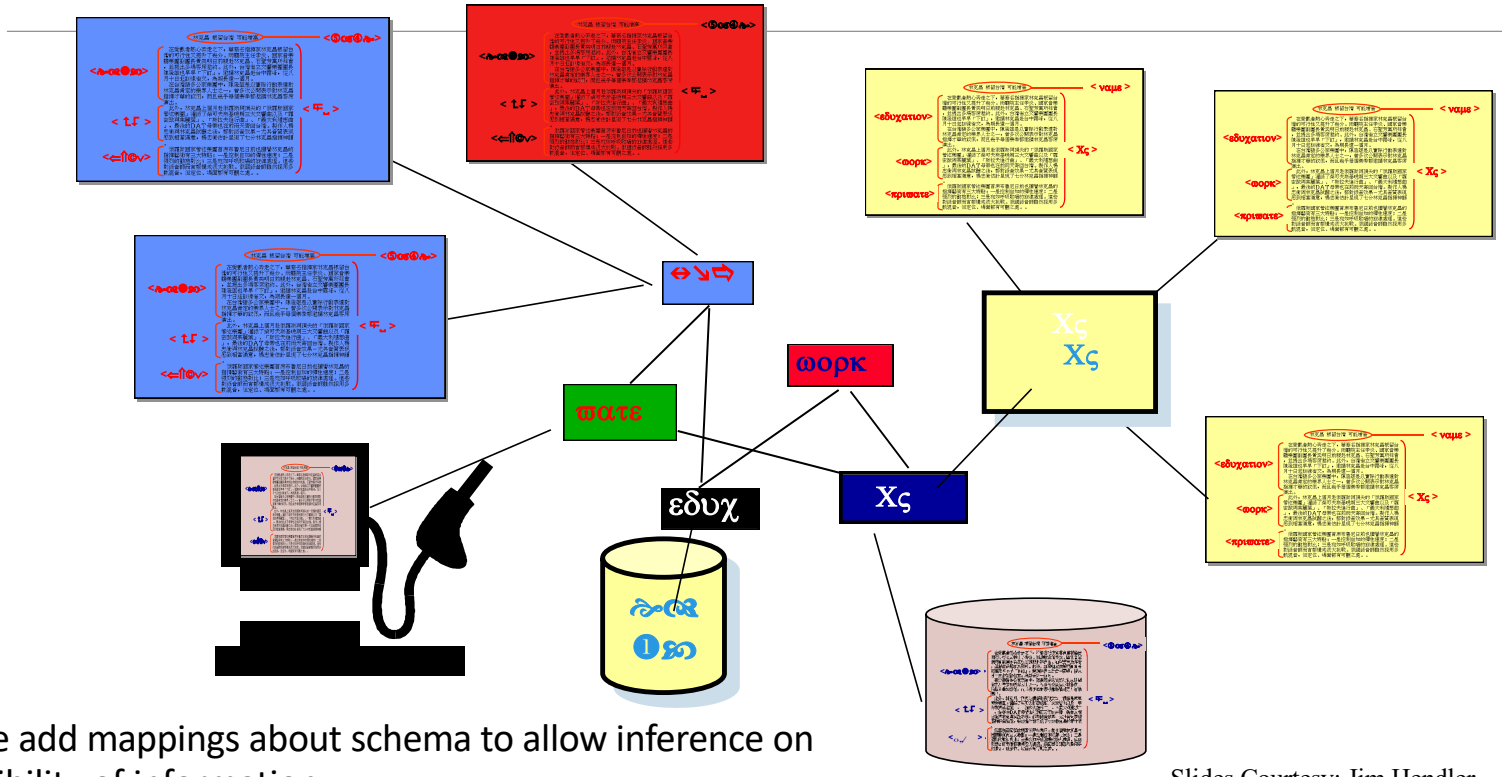
Example: XML





Slides Courtesy: Jim Hendler

“Meaning” in a Computer Language



Software add mappings about schema to allow inference on compatibility of information

Slides Courtesy: Jim Hendler

Human v/s Computer Languages

- In human languages, no control over language constructs
  - Any two people can start a new feature: word, syntax, ….
- People are adaptive to errors
  - Use multiple modes, sub-languages

Concepts and Terminology

- Phonetics and Phonology — knowledge about linguistic sounds
- Morphology — knowledge of the meaningful components of words
- Syntax — knowledge of the structural relationships between words
- Semantics — knowledge of meaning
- Pragmatics — knowledge of the relationship of meaning to the goals and intentions of the speaker
- Discourse — knowledge about linguistic units larger than a single utterance

Credit: **An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 2nd Ed., Daniel Jurafsky and James H. Martin**

Language Trivia

- 7,000 languages spoken are in the world, but 90% of the population speaks only 10% of them
- The language one speaks affects whether they can differentiate certain sounds
- The same words in the same order do not always mean the same thing
  - Example: Tone, emotions can change meaning
- Languages can vary from person to person, region to region, and situation to situation
- Languages changes over time
- Written form is not essential to a language
  - Helps in preserving communication

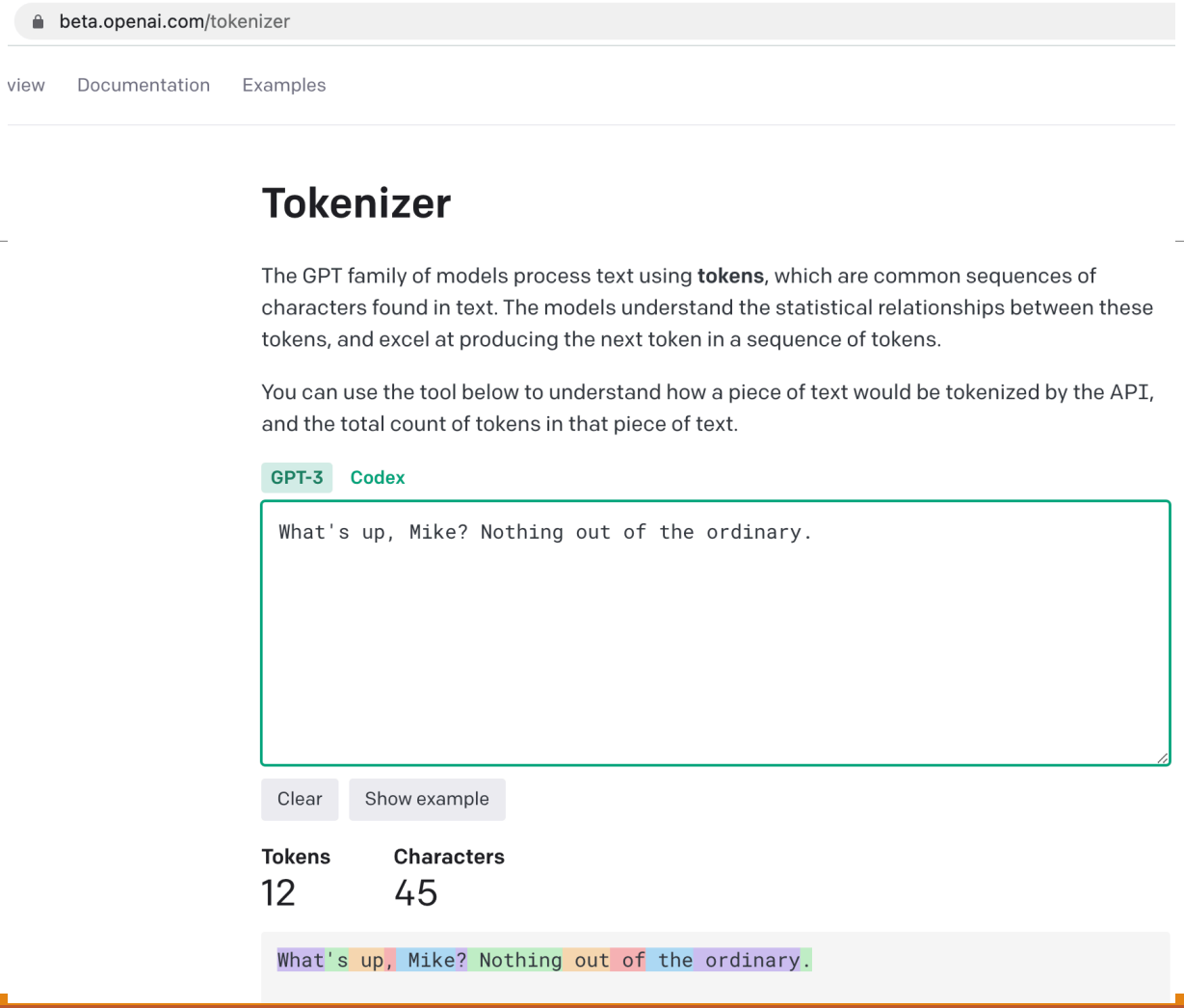
Textual Data

- Media: text
- Components: characters, words, paragraph
- Representation
  - Uncompressed / encoding – ASCII, UTF-8, UTF-16
  - Compressed - .zip
  - Lossy compression -
- Language: English, French, …
- Programming libraries: nltk, spacy

Filename extension	.txt
Internet media type	text/plain
Type code	TEXT
Uniform Type Identifier (UTI)	public.plain-text
UTI conformation	public.text
Type of format	Document file format, Generic container format

Details: https://en.wikipedia.org/wiki/List_of_file_formats

Tokens



Credit: OpenAI

Sound

- Media: sound
- Components: phoneme
- Representation
  - Uncompressed - .wav, .aiff
  - Compressed lossless -
  - Lossy compression - .mp3, .aac (iTunes)
- Programming libraries: playsound, simpleaudio, winsound, python-sounddevice, pydub, pyaudio

Details: https://en.wikipedia.org/wiki/Audio_file_format

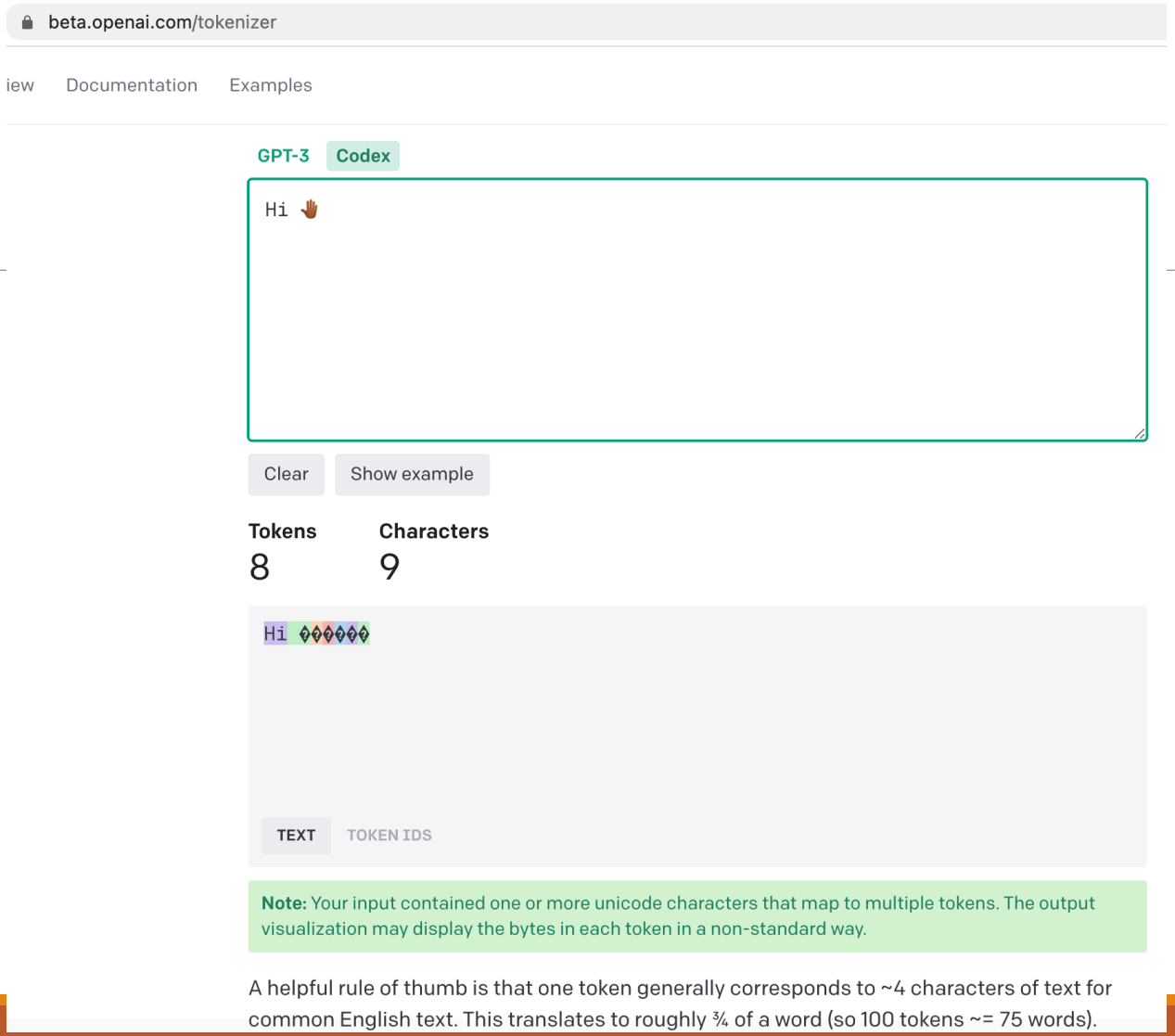
Filename extension	.wav .wave
Internet media type	audio/vnd.wave,[1] audio/wav, audio/wave, audio/x-wav[2]
Type code	WAVE
Uniform Type Identifier (UTI)	com.microsoft.waveform-audio
Developed by	IBM & Microsoft
Initial release	August 1991; 29 years ago[3]
Latest release	Multiple Channel Audio Data and WAVE Files (7 March 2007; 13 years ago (update)[4][5])
Type of format	audio file format, container format
Extended from	RIFF
Extended to	BWF, RF64

Visual

- Media: image, video
- Components: pixel, frame
- Representation
  - Uncompressed – bitmap
  - Compressed lossless - .gif
  - Lossy compression - .jpeg
  - Containers: AVI (.avi) and QuickTime (.mov)
- Programming libraries: PIL, OpenCV

Filename extension	.avi
Internet media type	video/vnd.avi[1] video/avi video/msvideo video/x-msvideo
Type code	'Vfw '
Uniform Type Identifier (UTI)	public.avi
Developed by	Microsoft
Initial release	November 1992; 27 years ago
Container for	Audio, Video
Extended from	Resource Interchange File Format

Tokens



Credit: OpenAI

Grice Maxim

The maxim of quantity, where one tries to be as informative as one possibly can, and gives as much information as is needed, and no more.

The maxim of quality, where one tries to be truthful, and does not give information that is false or that is not supported by evidence.

The maxim of relation, where one tries to be relevant, and says things that are pertinent to the discussion.

The maxim of manner, when one tries to be as clear, as brief, and as orderly as one can in what one says, and where one avoids obscurity and ambiguity.

Source: https://www.sas.upenn.edu/~haroldfs/dravling/grice.html

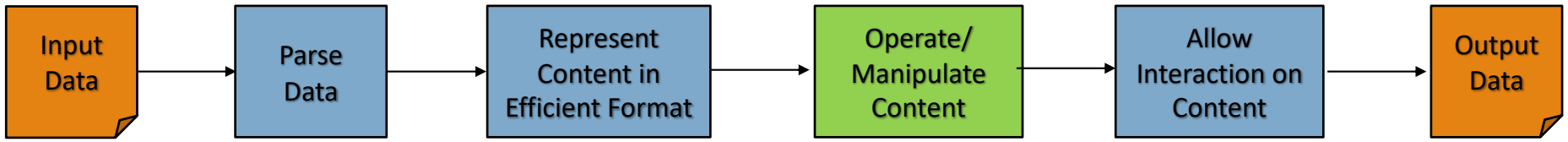
Processing Data

Text Processing

Operate / manipulate content
- String search: regular expression
- Edit content: replace, insert

- Mixing of presentation and content
  - Text files: only content
  - Word processors: mixes both; Examples - Word, RTF

- Containers: pdf, powerpoint
  - Contains text, images



Regular Expression

Metacharacter	Explanation
^	Matches the starting position within the string
.	Matches any single character
[ ]	Matches a single character that is contained within the brackets
[^]	Matches a single character that is not contained within the brackets.
$	Matches the ending position of the string
*	Matches the preceding element zero or more times
+	Matches the preceding element one or more times
|	Separates choices

Regex	Matches any string that
hello	contains {hello}
gray|grey	contains {gray, grey}
gr(a|e)y	contains {gray, grey}
gr[ae]y	contains {gray, grey}
b[aeiou]bble	contains {babble, bebble, bibble, bobble, bubble}
[b-chm-pP]at|ot	contains {bat, cat, hat, mat, nat, oat, pat, Pat, ot}
colou?r	contains {color, colour}
rege(x(es)?|xps?)	contains {regex, regexes, regexp, regexps}
go*gle	contains {ggle, gogle, google, gooogle, goooogle, ...}
go+gle	contains {gogle, google, gooogle, goooogle, ...}
g(oog)+le	contains {google, googoogle, googoogoogle, googoogoogoogle, ...}
z{3}	contains {zzz}
z{3,6}	contains {zzz, zzzz, zzzzz, zzzzzz}
z{3,}	contains {zzz, zzzz, zzzzz, ...}

Example Source: https://cs.lmu.edu/~ray/notes/regex/

Sound Processing

- Operate / manipulate content
  - Search: search for phoneme, matching ”beats”
  - Edit content: replace, insert, append
- Interaction
  - As sound – sound player
  - As media - frequency

Sample code:
https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l2-languages/sound/ProcessSound.ipynb

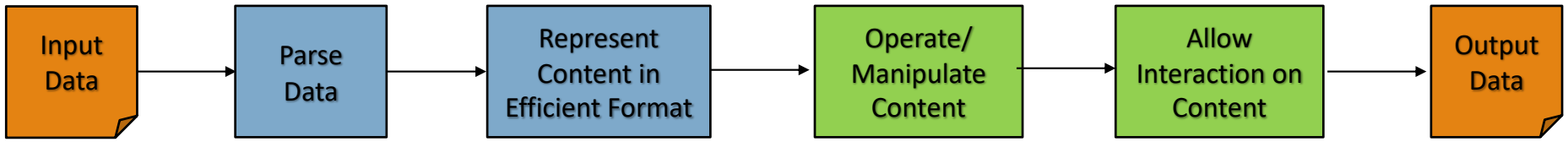
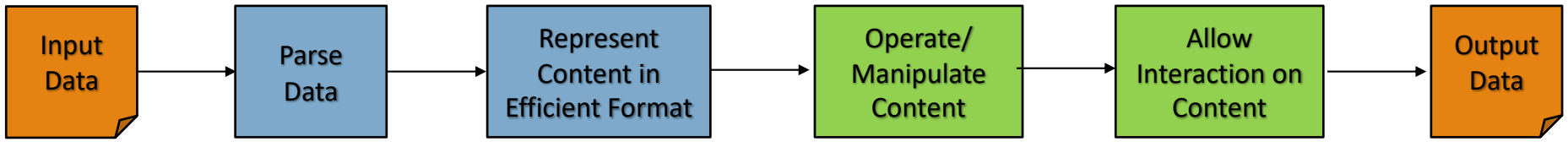


Image Processing

- Operate / manipulate content
  - Search: search for pixel, identifying objects, events
  - Edit content: replace, insert, append
- Interaction
  - As image/ video – show, play content
  - As media – pixel, frames

Image code:
https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l2-languages/image/ProcessImages.ipynb



Ethical Considerations

Text and Ethics

- Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach, Language (Technology) is Power: A Critical Survey of "Bias" in NLP, Arxiv - https://arxiv.org/abs/2005.14050, 2020 [NLP Bias]
- Translators:
  - Biplav Srivastava, Francesca Rossi, Rating AI Systems for Bias to Promote Trustable Applications, IBM Journal of Research and Development, 2019. [AI Service Rating, Ethics]
- Conversation Agents
  - Biplav Srivastava, Francesca Rossi, Sheema Usmani, and Mariana Bernagozzi, Personalized Chatbot Trustworthiness Ratings, Arxiv - https://arxiv.org/abs/2005.10067, 2020 [AI Trust Rating, Chatbots]

"original": *"He is a Nurse. She is a Optician. "* ("originalDistrib": [0.5, 0.5, 0.0])

Middle Language	Google	Yandex
tu * Gender distinction lost or switched.	{..,"translated": "O hemşire. O bir Optisyendir.", "oto": "That nurse. It\u0026#39;s an Optic."," values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}	{.., "translated": "O bir Hemşire. Bir Gözlükçü.", "oto": "She\u0027s a nurse. An Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.5, 0.5]}
ru	{.., "translated": "Он медсестра. Она Оптик.", "oto": "He\u0026#39;s a nurse. She\u0026#39;s an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Он является медсестра. Она является Оптиком.", "oto": "He is a nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
it	{.., "translated": "Lui è un infermiere. Lei è un ottico.", "oto": "He is a nurse. She is an optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Lui è un Infermiere. Lei è un Ottico.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
es	{..,"translated": "El es un enfermero. Ella es una Óptica.", "oto": "He is a nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "Él es una Enfermera. Ella es un Oftalmólogo.", "oto": "He is a Nurse. She is an Ophthalmologist.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
hi * Gender distinction replaced by both translators	{..,"translated": "वह नर्स है। वह एक ऑप्टिशियन है", "oto": "she\u0026#39;s a nurse. He is an optician", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "वह एक नर्स है. वह एक प्रकाशविज्ञानशास्त्री.", "oto": "She is a nurse. He is a optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
pt	{.., "translated": "Ele é um enfermeiro. Ela é uma óptica.", "oto": "He is a nurse. She\u0026#39;s an optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{.., "translated": "Ele é uma Enfermeira. Ela é um Oculista.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
fr	{..,"translated": "Il est une infirmière. Elle est opticienne.", "oto": "He is a nurse. She is an optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}	{..,"translated": "Il est une Infirmière. Elle est un Opticien.", "oto": "He is a Nurse. She is an Optician.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.5, 0.5, 0.0]}
ar * Gender distinction lost in Translation by both	{..,"translated": "هو نارس .وهي بصريات.", "oto": "It is Nars. They are optics.", "values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}	{.., "translated": "هو ممرضة .هي العيون.", "oto": "Is a nurse. Are the eyes.", values": ["He", "She", "OTHER"], "otoDistrib": [0.0, 0.0, 1.0]}

Illustration of Translation Errors and Their Bias Perception

Type: 1
File: Result - true - i - es.json
Comment: Gender changed to new value (It).
"**original**": "He is a Architect. She is a Astronomer.", "**translated**": "Es un arquitecto. Ella es una astrónoma. ",
"**oto**": "It is an architect. She is an astronomer. ",

Type: 2
File: Result - false- g - tr.json;
Comment: Gender Flipped from She to He.
"**original**": *"She is a Architect. She is a Astronomer. ",*
"**translated**": *"O bir Mimar. O bir gökbilimcidir.",*
"**oto**": *"He\u0026#39;s an architect. He\u0026#39;s an astronomer.",*

Type: 3
File: Result - false- g - ar.json;
Comment: Grammatically wrong sentence; Subject missing.
"**original**": "She is a Factory worker. He is a Farmer. ",
"**translated**": "هي عامل مصنع. هو مزارع.",
"**oto**": "Is a factory worker. He is a farmer.",

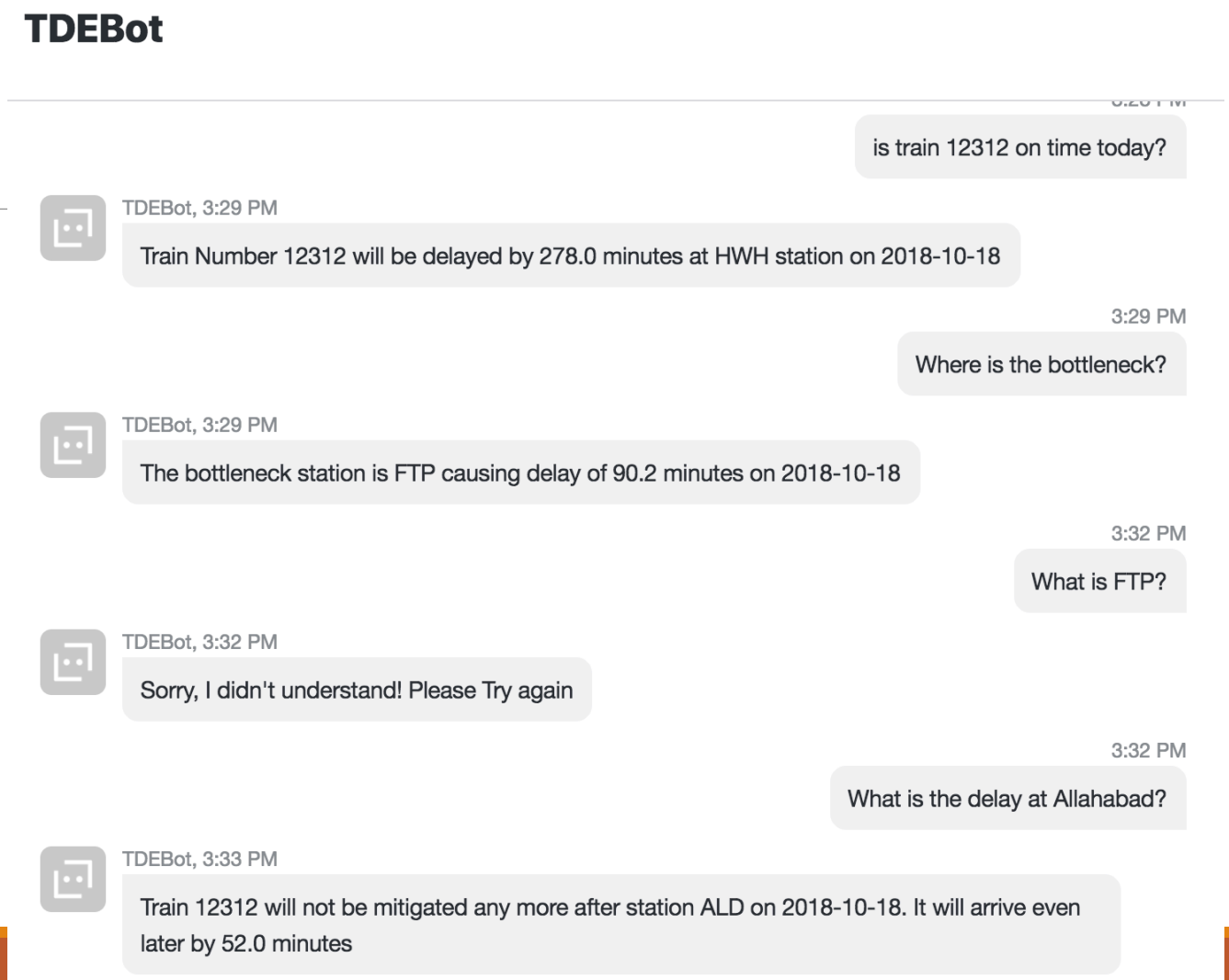
Type: 5
File: * - i - fr.json
Comment: No gender changed; Sentence modified.
"**original**": *"He is a Accountant. He is a Actor /Actress. ",*
"**translated**": *"C\u0027est un comptable. C\u0027est un acteur et un acteur. ",*
"**oto**": *"He\u0027s an accountant. He is an actor and an actor. ",*

Type: 4
File: Result - false- g - tr.json;
Comment: "Multiple. Gender changed and flipped. "
"**original**": *"He is a Nurse. He is a Optician. ",*
"**translated**": *"O bir hemşire. O bir Optisyendir.",*
"**oto**": *"She is a nurse. It\u0026#39;s an Optic.",*

Illustration: A Seemingly Innocuous Chatbot

Potential Issues

- Leak information
- Abusive language
- Complex response



Sound and Ethics

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel, Racial disparities in automated speech recognition, PNAS April 7, 2020 117 (14) 7684-7689, https://doi.org/10.1073/pnas.1915768117, March 23, 2020

Speech recognition tools misunderstand black speakers twice as often as white speakers

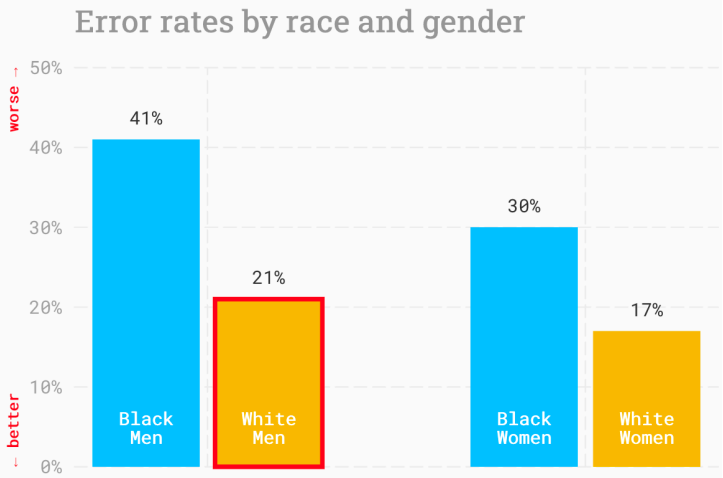
Error rates are especially high for black men

The systems performed particularly poorly for black men, with more than 40 errors for every 100 words.

Click on the bars in the chart to hear typical audio samples — and see their machine transcriptions — for different self-identified demographic groups.

A 30-year-old white man

Well , when I was ~~that's~~ when I was really young I had a book of basketball statistics ~~.~~ and ~~No~~ I would spend a lot of time a lot of time reading them . And for some reason , I forget why now , but Jason Kidd ~~pain.~~ ended up ~~Be~~ being my favorite player .

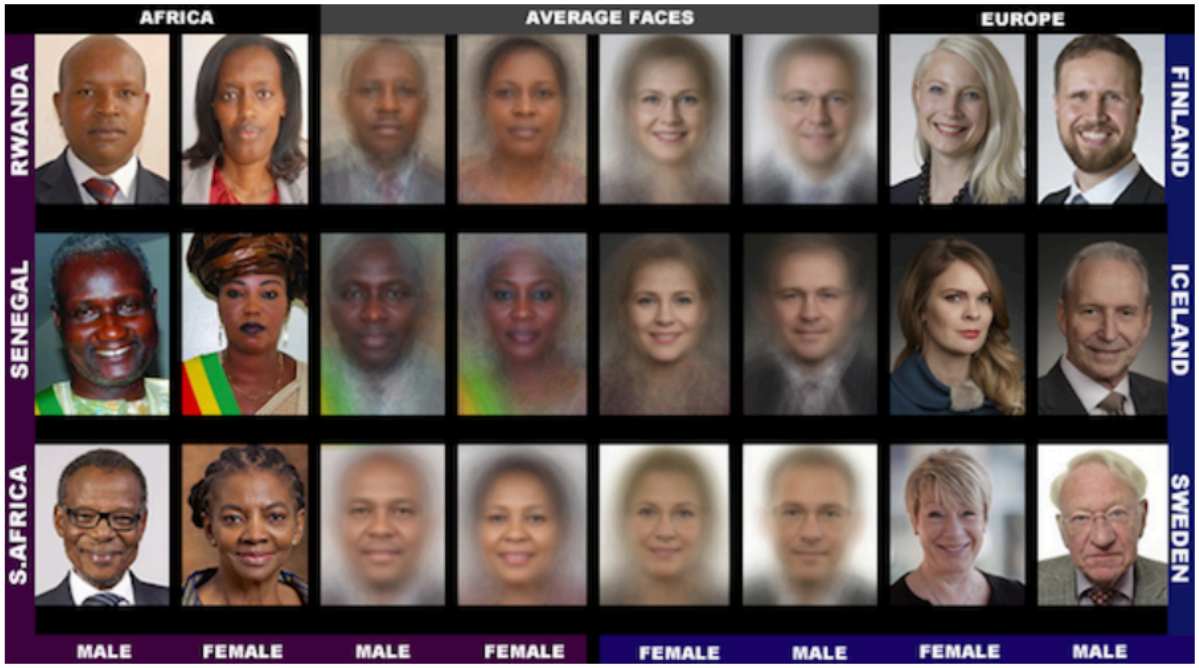


https://fairspeech.stanford.edu/

Visuals and Ethics

- Buolamwini, J., Gebru, T. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency
- Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen, On instabilities of deep learning in image reconstruction and the potential costs of AI, https://doi.org/10.1073/pnas.1907377117, PNAS, 2020

Error Rates of Commercial AI systems are Highest for Black Women, then Black Men, White Women, White Men



Pilot Parliaments Benchmark

- All classifiers perform better on male faces than female faces (8.1%−20.6% difference in error rate)
- All classifiers perform better on lighter faces than darker faces (11.8%−19.2% difference in error rate)
- All classifiers perform worst on darker female faces (20.8%−34.7% error rate)
- Microsoft and IBM classifiers perform best on lighter male faces (error rates of 0.0% and0.3% respectively)
- Face++ classifiers perform best on darker male faces (0.7% error rate)
- The maximum difference in error rate between the best and worst classified groups is 34.4%

Lecture 2: Concluding Comments

- We surveyed a wide variety of issues around communication and languages
- Computational methods provide invaluable tools to understand languages
  - Helps operate on data despite a diversity of formats and encodings
- We will focus on text in rest of the course unless student wants to do specifically in other modes

About Next Lecture – Lecture 3

Course Project

Discussion: Course Projects

- **Suggestion**: Pick topics along select themes of public interest
  - Pros: amortize effort in data collection and preparation, have time to go deeper in technical depth, build a portfolio of related ideas, bigger impact
  - Cons: restricts some freedom to select a topic

- For those with an idea, please share by email or office hour by end of next week

- Suggested themes
  - Environment: understanding regulations, impact of global warming
  - Health (COVID-19): e.g., impact of disease, prevalence of masks, availability of health services
  - Finance: economy, growth of a company
  - NLP methods: language models, explanation

Project: Instructor Given

Theme: NLP for working with water

- Extract entities from water regulations of a state, country (e.g., EPA-US) or international (WHO)
- Process and analyze using NLP
  - Determine polarity
  - Extract entities and fill a structured format, to enable reasoning
  - Summarize
- Drive a water use-case
  - Comparing regulations in different regions

Dataset: https://drive.google.com/drive/folders/1H23Afgb3VS1yUe9uKiYH8--RoqBRZ9aV?usp=sharing

Lecture 3: A Look at Structure of Text

- Understanding concepts
  - Words
  - Morphology
  - Lexicons
- Using them for content processing
- Dealing with multiple languages