# Special Topics in Natural Language Processing

## CS698O

Ashutosh Modi
CSE Department, IIT Kanpur


Lecture 2: Why NLP is Hard and Linguistic Fundamentals-1
Jan 6, 2020

# Why language is difficult to process *Computationally*?

# Language is Not an Array of Words

# Language is Not an Array of Words

- Meaning of a sentence can completely change with change in just one word

# Language is Not an Array of Words

- Meaning of a sentence can completely change with change in just one word

*Every Indian has a mother.*

*vs*

*Every Indian has a President*.

# Language is Not an Array of Words

- Meaning of a sentence can completely change with change in just one word

*Every Indian has a mother.*

*vs*

*Every Indian has a President*.

*We gave monkey bananas because they were hungry.*

*vs*

*We gave monkey bananas because they were over-ripe*.

Examples: [1]

# Language is Not an Array of Words

- Word order matters

*The pen is in the box.*

*vs*

*The box is in the pen.*

Examples: [1]

# Language is Not an Array of Words

- Word order matters

*The pen is in the box.*

*vs*

*The box is in the pen.*

*John gave Mary a gift on birthday.*

*vs*

*Mary gave John a gift on birthday.*

# Ambiguity

- Language is Ambiguous
- Ambiguity makes communication efficient, but this makes computational processing challenging

# Ambiguity

## Lexical Ambiguity

*He went to the bank and saw a lot of people there.*


vs.


Image: Source Unknown. CC license

# Ambiguity

## Syntactic Ambiguity

*He saw a man with binoculars.*

**vs.**

Image: Source Unknown. CC license

# Ambiguity

## Semantic Ambiguity

*Rachel paid the waitress and then she left the restaurant.*

**vs.**

Image: https://www.eatthis.com/

# Infinite Constructions

- Paraphrases: Same concept can be expressed in multiple different ways.

*Microsoft purchased the Canadian company Maluuba.*

*A.I. company Maluuba was bought by Microsoft on January 9, 2017.*

*Microsoft purchased Maluuba for $30 million.*

*In a recent move, Microsoft bought the Canadian startup Maluuba.*

# Syntax vs Semantics

- A sentence may be syntactically correct but may not make any sense.

# Syntax vs Semantics

- A sentence may be syntactically correct but may not make any sense.

*Colorless green ideas sleep furiously*

Chomsky 1957

# Ambiguity can be problematic

*Astronaut Takes Blame for Gas in Spacecraft*

*March Planned For Next August*

Source: [2]

# Ambiguity can be problematic

*Astronaut Takes Blame for Gas in Spacecraft*

*March Planned For Next August*

*Aging Expert Joins University Faculty*

*Child teaching expert to speak*

*Two Sisters Reunited After 18 Years in Checkout Counter*

Source: [2]

# Ambiguity can be problematic

*Astronaut Takes Blame for Gas in Spacecraft*

*March Planned For Next August*

*Aging Expert Joins University Faculty*

*Child teaching expert to speak*

*Two Sisters Reunited After 18 Years in Checkout Counter*

*Deaf College Opens Doors to Hearing*

Source: [2]

# Language Evolves

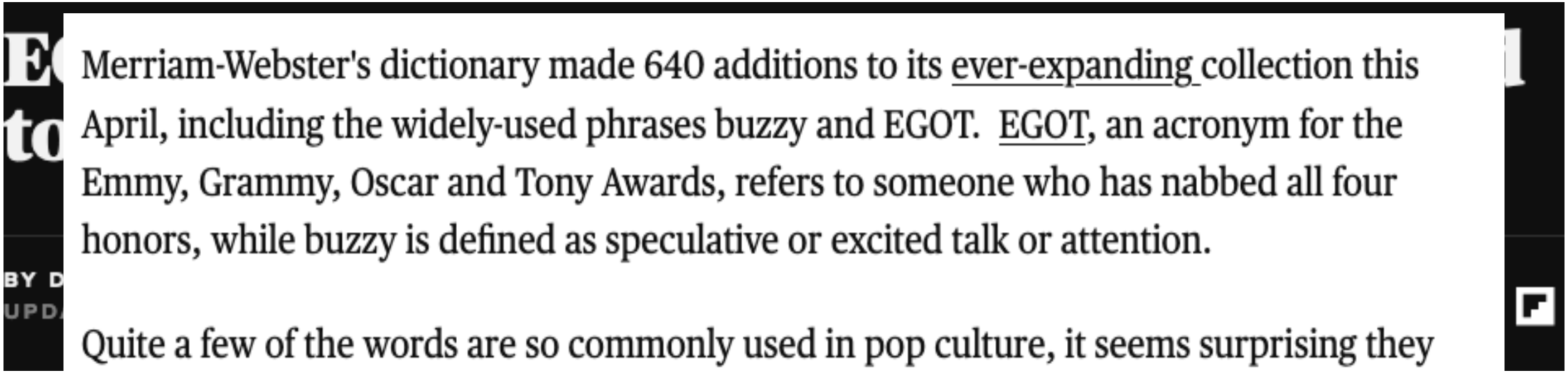
**EGOT, buzzy among 640 new words added to Merriam-Webster dictionary**

BY DANIELLE GARRAND

UPDATED ON: APRIL 24, 2019 / 11:47 AM / CBS NEWS

Source: https://www.cbsnews.com/news/merriam-webster-dictionary-new-words-added-to-dictionary-april-2019-buzzy-egot-among-640-new-words-added/

# Language Evolves



Merriam-Webster's dictionary made 640 additions to its ever-expanding collection this April, including the widely-used phrases buzzy and EGOT. EGOT, an acronym for the Emmy, Grammy, Oscar and Tony Awards, refers to someone who has nabbed all four honors, while buzzy is defined as speculative or excited talk or attention.

Quite a few of the words are so commonly used in pop culture, it seems surprising they haven't been added to the dictionary yet.

Also added to the lexicon: The millennial favorites stan, a very devoted fan; peak, now also defined as being at the height of popularity, use, or attention; and on-brand, to be consistent with an image or identity.

Source: https://www.cbsnews.com/news/merriam-webster-dictionary-new-words-added-to-dictionary-april-2019-buzzy-egot-among-640-new-words-added/

# Language evolves

- Non-standard usage of words

  *friend as noun vs  friend as verb*

  *spam used to be food but now it used differently*

# Language evolves

- Non-standard usage of words

  *friend as noun vs  friend as verb*

  *spam used to be food but now it used differently*

- Slangs

  *Selfie, Chillax*

# Language evolves

- Non-standard usage of words

  *friend as noun vs  friend as verb*

  *spam used to be food but now it used differently*

- Slangs

  *Selfie, Chillax*

- Code mixing

  Hindi + English = Hinglish

  *ICON 2016 Varanasi me hold hoga! Great chance to see the pracheen nagari!*

  (ICON 2016 will be held in Varanasi! Great chance to see the ancient city!)

# Language evolves

- Non-standard usage of words

  *friend as noun vs  friend as verb*

  *spam used to be food but now it used differently*

- Slangs

  *Selfie, Chillax*

- Code mixing

  Hindi + English = Hinglish

  *ICON 2016 Varanasi me hold hoga! Great chance to see the pracheen nagari!*

  (ICON 2016 will be held in Varanasi! Great chance to see the ancient city!)

- Social Media Text

  Twitter with hashtags, different spellings, etc.

# Language is challenging

- Metaphors

*Time flies like an arrow.*

# Language is challenging

- Metaphors

*Time flies like an arrow.*

*It is raining cats and dogs.*

# Language is challenging

- Sarcasm

*Movie was so awesome that I left after half an hour.*

# Language is challenging

- Poetry

:
:
*है कौन विघ्न ऐसा जग में,*
*टिक सके वीर नर के मग में*
*खम ठोंक ठेलता है जब नर,*
*पर्वत के जाते पाँव उखड़।*
*मानव जब जोर लगाता है,*
*पत्थर पानी बन जाता है।*
:

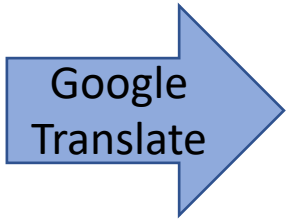
*Rashmi-Rathi, part 3, Ramdhari Singh 'Dinkar'*

Poem Link: https://www.hindi-kavita.com/HindiRashmirathiDinkar.php#Rashmirathi3

# Language is challenging

- Poetry

:
:
*है कौन विघ्न ऐसा जग में,*
*टिक सके वीर नर के मग में*
*खम ठोंक ठेलता है जब नर,*
*पर्वत के जाते पाँव उखड़।*
*मानव जब जोर लगाता है,*
*पत्थर पानी बन जाता है।*
:




:
:
Who is disturbed in this world,
Could stand in the mug of a brave man
Shatters when the male,
Feet dislocated while going to the mountain.
When a human exerts force,
The stone becomes water.
:

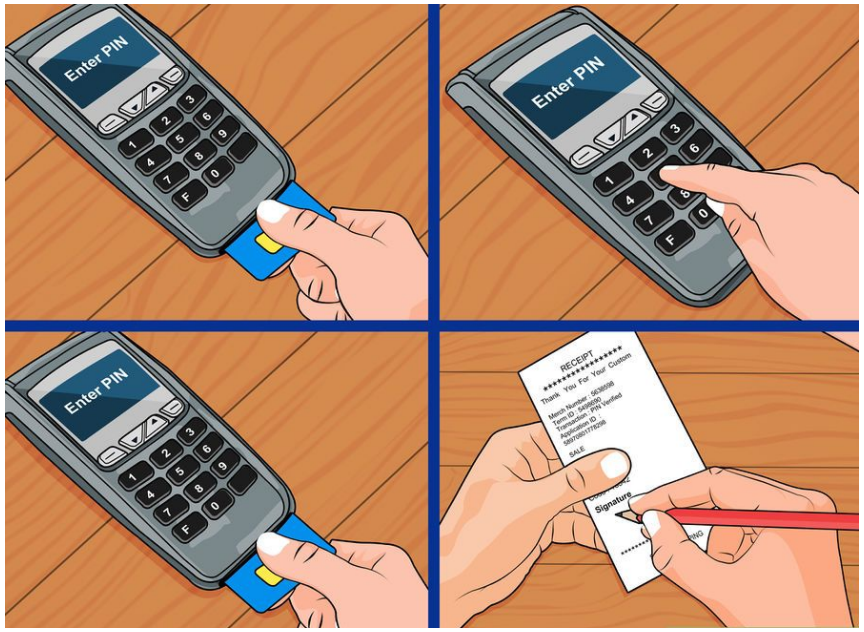
*Rashmi-Rathi, part 3, Ramdhari Singh 'Dinkar'*

Poem Link: https://www.hindi-kavita.com/HindiRashmirathiDinkar.php#Rashmirathi3

# Language is challenging

- Language is implicitly grounded in world knowledge

*I paid the bill and left.*




Example: [3]

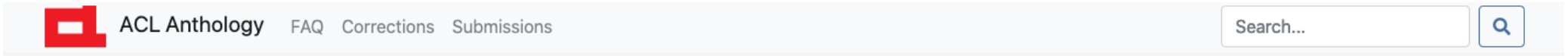
*Image: https://www.wikihow.life/Use-a-Debit-Card*

# NLP Conferences and Research Resource

- ACL (Association of Computational Linguistics)
- EMNLP (Empirical Methods in NLP)
- CoNLL (Computational Natural Language Learning)
- CoLing (Computational Linguistics)
- NAACL
- EACL
- TACL (Journal: Transactions of ACL)

https://www.aclweb.org/anthology/

# https://www.aclweb.org/anthology/




## Welcome to the ACL Anthology!

The ACL Anthology currently hosts 53854 papers on the study of computational linguistics and natural language processing.

Subscribe to the mailing list to receive announcements and updates to the Anthology.

The Anthology can archive your poster or presentation! Please submit them in PDF format by **filling out this form.** Attachments will be distributed under the terms of the **CC-BY-4.0 license.**

Full Anthology as BibTeX

### ACL Events

| Venue | Present – 2010 | 2009 – 2000 | 1999 – 1990 | 1989 and older |
|---|---|---|---|---|
| ACL | 19 18 17 16 15 14 13 12 11 10 | 09 08 07 06 05 04 03 02 01 00 | 99 98 97 96 95 94 93 92 91 90 | 89 88 87 86 85 84 83 82 81 80 79 |
| ANLP | | 00 | 97 94 92 | 88 83 |
| CL | 19 18 17 16 15 14 13 12 11 10 | 09 08 07 06 05 04 03 02 01 00 | 99 98 97 96 95 94 93 92 91 90 | 89 88 87 86 85 84 83 82 81 80 78 77 76 |
| CoNLL | 19 18 17 16 15 14 13 12 11 10 | 09 08 07 06 05 04 03 02 01 00 | 99 98 97 | |
| EACL | 17 14 12 | 09 06 03 | 99 97 95 93 91 | 89 87 85 83 |
| EMNLP | 19 18 17 16 15 14 13 12 11 10 | 09 08 07 06 05 04 03 02 01 00 | 99 98 97 96 | |
| NAACL | 19 18 16 15 13 12 10 | 09 07 06 04 03 01 00 | | |
| *SEMEVAL | 19 18 17 16 15 14 13 12 10 | 07 04 01 | 98 | |
| TACL | 19 18 17 16 15 14 13 | | | |
| WS | 19 18 17 16 15 14 13 12 11 10 | 09 08 07 06 05 04 03 02 01 00 | 99 98 97 96 95 94 93 92 91 90 | 89 88 86 84 81 79 77 |
| SIGs | ANN \| BIOMED \| DAT \| DIAL \| EDU \| FSM \| GEN \| HAN \| HUM \| LEX \| MEDIA \| MOL \| MORPHON \| MT \| NLL \| PARSE \| REP \| SEM \| SEMITIC \| SLAV \| SLPAT \| UR \| WAC | | | |

### Non-ACL Events

| Venue | Present – 2010 | 2009 – 2000 | 1999 – 1990 | 1989 and older |
|---|---|---|---|---|
| ALTA | 19 18 17 16 15 14 13 12 11 10 | 09 08 07 06 05 04 03 | | |
| COLING | 18 16 14 12 10 | 08 06 04 02 00 | 98 96 94 92 90 | 88 86 82 80 73 69 67 65 |
| HLT | | 06 05 04 03 01 | 94 93 92 91 90 | 89 86 |
| IJCNLP | 19 17 15 13 11 | 09 08 05 | | |
| JEP/TALN/RECITAL | 14 13 12 | | | |
| LREC | 18 16 14 12 10 | 08 06 04 02 00 | | |
| MUC | | | 98 95 93 92 91 | |
| PACLIC | 18 17 16 15 14 13 12 11 10 | 09 08 07 06 05 04 03 01 00 | 99 98 96 95 | |
| RANLP | 17 15 13 11 | 09 | | |
| ROCLING/IJCLCLP | 18 17 16 15 14 13 12 11 10 | 09 08 07 06 05 04 03 02 01 00 | 99 98 97 96 95 94 93 92 91 90 | 89 88 |
| TINLAP | | | | 87 78 75 |
| TIPSTER | | | 98 96 93 | |

Special To

# LINGUISTIC FUNDAMENTALS

# Linguistic Fundamentals

- Reviews basic structure of English in terms of word categories and phrases categories

- It applies to other languages as well

Note: Most of this section is based on Chapter 2 of [4]

# Words

- “Word” appears to be the most basic unit of a language

*Running* *Drank*

# Words

- "Word" appears to be the most basic unit of a language

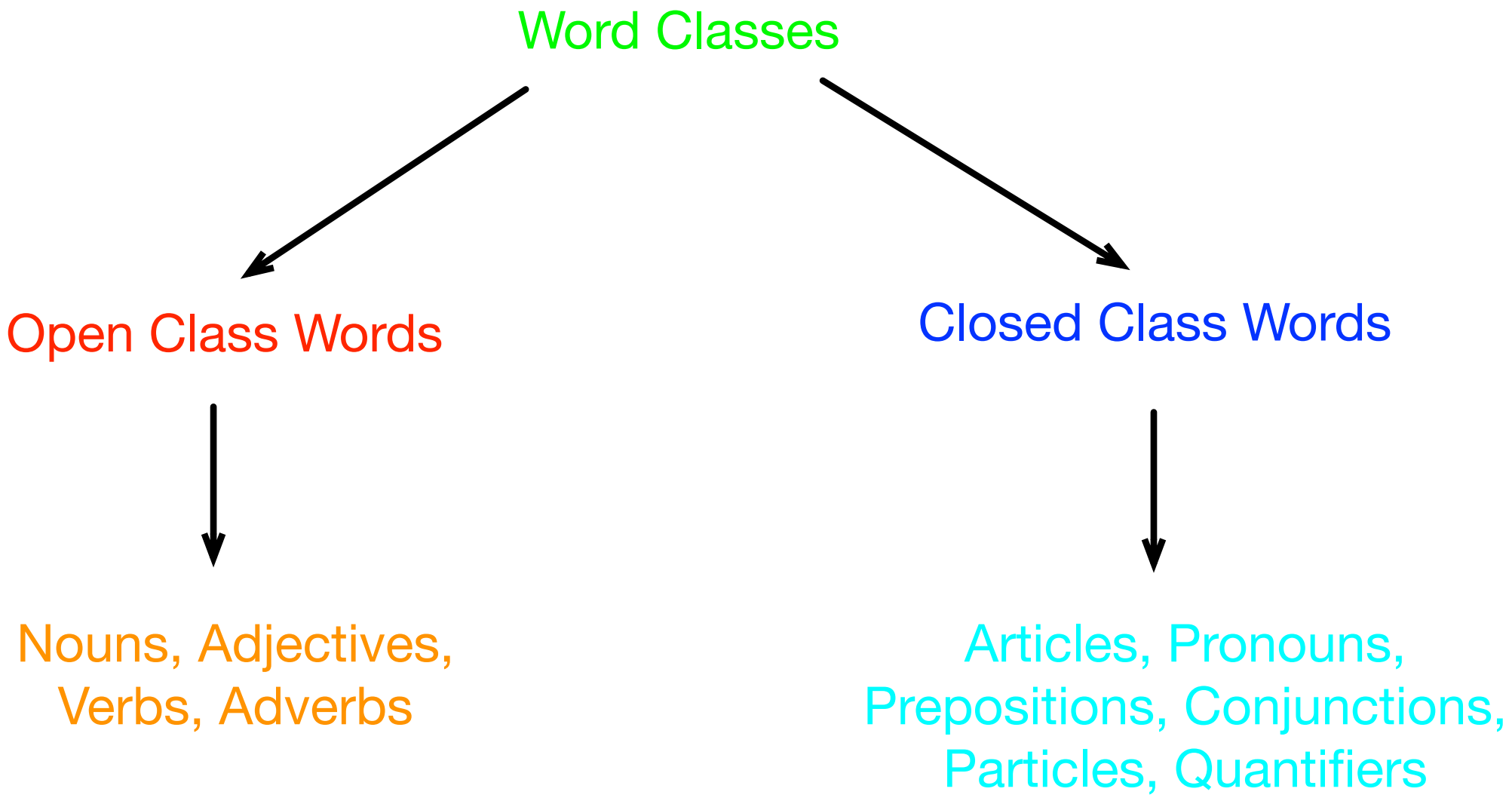
*Running* *Drank*

- But word are result of complex set of rules on more primitive parts.

- Morphology is the study of how words are built up from more basic components (**Morphemes**) corresponding to minimal meaning units.

*Run → Ran, Running*

*Goose → Geese* *Fox → Foxes*


Word Classes
Open Class Words
Closed Class Words
Nouns, Adjectives,
Verbs, Adverbs
Articles, Pronouns,
Prepositions, Conjunctions,
Particles, Quantifiers

# Phrases

- Phrase is one or more words functioning as a unit in sentence. [5]
- Every phrase has a head
- Head indicates the type of thing, action or quality that phrase describes. Head belongs to open class words

*the dog*

*the ugly dog*

*the ugly dog at the ground*

# Phrases

- Noun Phrases
- Adjective Phrases
- Verb Phrases
- Adverbial Phrases

*The desire to succeed*

*angry as a hippo*

*ate the cake*

*rapidly like a bird*

# Phrases

- Many times a phrase needs additional phrases following it to express desired meaning

*Jack put*

# Phrases

- Many times a phrase needs additional phrases following it to express desired meaning

*Jack put*

- These additional phrase(s) needed to complete the meaning are called the **complement of the head**

*Jack put the dog in the house*

# Noun Phrases (NP)

- NP refer to objects, places, concepts, events, qualities, etc.
- Head of NP is usually a common noun
- In some cases, NP can have a pronoun as the head

*It* hid under the rug
Once I opened the door, I regretted *it* for months

- In some cases, NP can have proper noun or count noun or mass nouns as head

*John* at food
*Dogs* are friendly
*Water* is necessary for life

# Noun Phrases (NP)

- NP may contain ***Specifiers*** and ***Qualifiers*** preceding the head
- Qualifiers describe the general class of objects identified by the head
- Specifiers indicate how many such objects are being described

**Specifier Qualifier Head**

*the ceiling paint can*

*the angry bird*

# Noun Phrases (NP)

- Specifiers constructed out of:
  - Ordinals (*first, second*, etc.)
  - Cardinals (*one, two*, etc.)
  - Determiners
    - Articles (*the, a*, and *an)*
    - Demonstratives (*this, that, these, those*, etc.)
    - Possessives (John*'s*, the fat man*'s*)
    - Wh-Determiners (question related word e.g. *which* and *what*)
    - Quantifying Determiners (*some, every, most, any, half*, etc.)

*the first three contestants*

# Summary

- NLP is hard
- Language is complex and inherently ambiguous
- It is good to know some linguistic fundamentals to develop computational models
- Words are divided into open classes and closed classes
- Based on open classes we have different phrase types.

# References


1. Dragomir Radev's Lecture Video: https://youtu.be/NHtohvD7gxY
2. Funny Headlines: https://www.ling.upenn.edu/~beatrice/humor/headlines.html
3. Modeling Common Sense Knowledge via Scripts: https://tinyurl.com/w2j2c29
4. Natural Language Understanding, James Allen (Chapter 2)
5. http://www.ello.uos.de/field.php/Syntax/SentenceClauseAndPhrase

- Next class:
  - Linguistic Fundamentals continued
  - NLP tools of trade