# CS689: Computational Linguistics for Indian Languages
# Word Processing

Arnab Bhattacharya
arnabb@cse.iitk.ac.in

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
http://web.cse.iitk.ac.in/~cs689/

2$^{nd}$ semester, 2023-24

Tue 10:30–11:45, Thu 12:00–13:15 at RM101/KD102

# Language

- Language is a particular encoding of communication between speaker/writer and listener/reader
- Representation of *mental* form of *idea*
- Natural language is complex due to ambiguity, multiple word meanings, context, idiomatic usage, sarcasm, intented meaning, etc.
- How does one *learn* a language?
- Rationalist view: human brains are hardwired with innate language capability
  - Rules of language (grammar, etc.) are learnt gradually
  - This fine-tunes language usage capabilities
- Empiricist view: language is learnt by exposure to usage
  - Pattern recognition, association, and generalization happen gradually
  - Statistical NLP

# Studies

- Knowing a language involves knowing several studies
- Phonetics and Phonology: sound
- Morphology: words and its parts
- Syntax: structural relationships among words
- Semantics: meanings
- Pragmatics: relationships of meanings to goals and intentions of writer/speaker
- Discourse: organization or flow of knowledge and thought

# Characters and Phonemes

- Characters are the basic units for *written* language
- Phonemes are the basic units for *spoken* language
- Indian languages have a one-to-one correspondence between characters and phonemes
  - Sanskrit has it perfectly, while other Indian languages follow it mostly
  - English and other languages
- English pronunication arbitrariness
  - Compare with “but” versus “put”: ‘u’ sounds differently
  - Characters ‘c’ and ‘k’ sound the same: “cake”
  - Characters not pronounced: ‘e’ in “cake”
  - Double characters pronounced only once: ‘l’ in “will”
- “What You See Is What You Pronounce” (WYSIWYP) for Indian languages
  - Elaborate and scientific *varna* system
  - $5 \times 5$ for the varga varṇas
  - Rows for different pronunciation locations
  - Columns for voiced/unvoiced, aspirated/unaspirated, nasal/non-nasal distinctions

# “Euro-English”

In the first year, ”s” will replace the soft ”c”. Sertainly, this will make the sivil servants jump with joy. The hard ”c” will be dropped in favour of the ”k”. This should klear up konfusion and keyboards kan have 1 less letter.
There will be growing publik enthusiasm in the sekond year, when the troublesome ”ph” will be replaced with ”f”. This will make words like ”fotograf” 20% shorter.
In the 3rd year, publik akseptanse of the new spelling kan be ekspekted to reach the stage where more komplikated changes are possible. Governments will enkorage the removal of double letters, which have always ben a deterent to akurate speling. Also, al wil agre that the horible mes of the silent ”e”s in the language is disgraseful, and they should go away.
By the fourth year, peopl wil be reseptiv to steps such as replasing ”th” with ”z” and ”w” with ”v”. During ze fifz year, ze unesesary ”o” kan be dropd from vords kontaining ”ou” and similar changes vud of kors be aplid to ozer kombinations of leters.

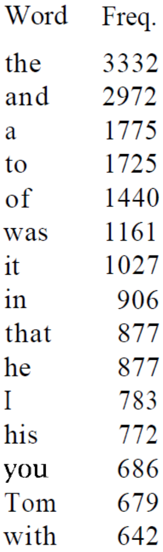
After zis fifz yer, ve vil hav a reli sensibl riten styl. Zer vil be no mor trubl or difikultis and evrivun vil find it ezi to understand ech ozer. Ze drem vil finali kum tru! And zen world!

# Words

- Words are combinations of characters that have a meaning *without* whitespaces (hyphens are allowed)
  - Any lesser unit changes the meaning
  - Whitespaces separate words
- Tokens are semantically meaningful combinations of characters
  - May be a single word or a sequence of words
  - "Uttar Pradesh"
  - "`Rāma se`" (from Rama)
- Words are *typographical* conventions while tokens are *semantic* units
- Tokenization is the process of breaking a text into tokens
- Vocabulary is the count of *unique* tokens
  - Mostly measured for a corpus
  - Sanskrit as a language has an infinite vocabulary

# Word Counts

- How are word counts in a language (corpus) distributed?



| Word | Freq. |
|---|---|
| the | 3332 |
| and | 2972 |
| a | 1775 |
| to | 1725 |
| of | 1440 |
| was | 1161 |
| it | 1027 |
| in | 906 |
| that | 877 |
| he | 877 |
| I | 783 |
| his | 772 |
| you | 686 |
| Tom | 679 |
| with | 642 |

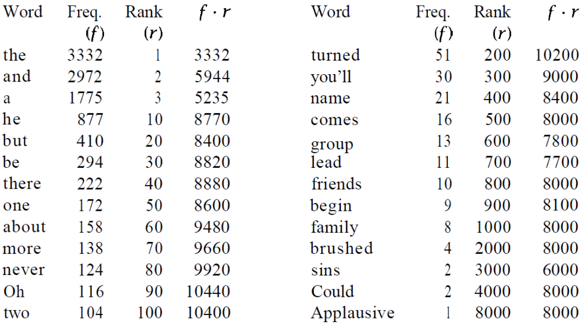
- Corpus is *Tom Sawyer* by Mark Twain (total: 71,370)

# Zipf's Law

- Frequency of a word is $f$
- *Rank* of a word according to frequency is $r$
- Zipf's law states that

$$f \propto 1/r \text{ or, } f \cdot r = \text{constant}$$

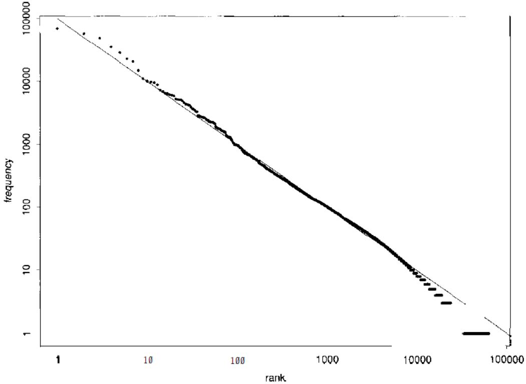
- For *Tom Sawyer* corpus



| Word | Freq. | Rank | $f \cdot r$ | Word | Freq. | Rank | $f \cdot r$ |
|---|---|---|---|---|---|---|---|
| | ($f$) | ($r$) | | | ($f$) | ($r$) | |
| the | 3332 | 1 | 3332 | turned | 51 | 200 | 10200 |
| and | 2972 | 2 | 5944 | you'll | 30 | 300 | 9000 |
| a | 1775 | 3 | 5235 | name | 21 | 400 | 8400 |
| he | 877 | 10 | 8770 | comes | 16 | 500 | 8000 |
| but | 410 | 20 | 8400 | group | 13 | 600 | 7800 |
| be | 294 | 30 | 8820 | lead | 11 | 700 | 7700 |
| there | 222 | 40 | 8880 | friends | 10 | 800 | 8000 |
| one | 172 | 50 | 8600 | begin | 9 | 900 | 8100 |
| about | 158 | 60 | 9480 | family | 8 | 1000 | 8000 |
| more | 138 | 70 | 9660 | brushed | 4 | 2000 | 8000 |
| never | 124 | 80 | 9920 | sins | 2 | 3000 | 6000 |
| Oh | 116 | 90 | 10440 | Could | 2 | 4000 | 8000 |
| two | 104 | 100 | 10400 | Applausive | 1 | 8000 | 8000 |

# Empirical Evaluation

- Empirical evaluation of Zipf's law on *Brown Corpus* (million words)



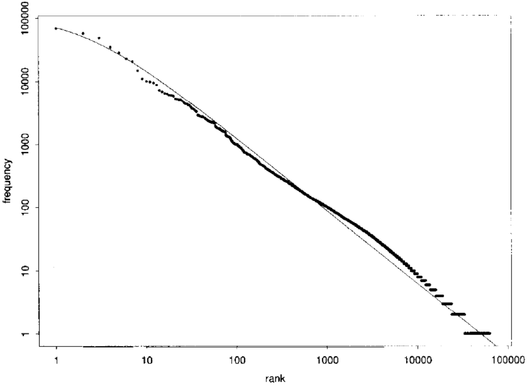
- Mostly all right, except for very high and very low ranks

# Mandelbrot's Formula

- Mandelbrot's formula

$$f = P \cdot (r + \rho)^{-B}$$

- Generalization of Zipf's law (where $B = 1$, $\rho = 0$)



- $\rho = 100$, $B = 1.15$, $P = 10^{5.4}$

# Word Meanings

- Every word has *at least* one meaning
- A word may have multiple meanings: homonym
- Multiple words may have "same" meaning: synonym
- Grossly same meaning, but may have nuances
    - vahniḥ (वह्निः), jvalanaḥ (ज्वलनः), pāvakaḥ (पावकः), śuṣmā (शुष्मा) all are synonyms of agniḥ (अग्निः) or "fire"
        - vahniḥ (वह्निः): that carries
        - jvalanaḥ (ज्वलनः): that burns
        - pāvakaḥ (पावकः): that purifies
        - śuṣmā (शुष्मा): that dries
        - agniḥ (अग्निः): that goes forward (upward)

# Number of Word Meanings

- Does a word with more meanings gets used more or less?
- Speaker wants words with more and more meanings
- Listener wants words with less and less meanings
- Zipf's not-so-famous law states that number of meanings of a word

$$m \propto \sqrt{f}$$

  - Prepositions, articles, conjunctions, etc. are not considered in English
- *Word vector* approach
  - Vectors can be added, subtracted, found similarity
  - A word is known by the company it keeps
  - Can a vector "predict" surrounding words?

# Types of Word Meanings

- Essential nature of a word or pada is śakti or vṛtti (roughly, power of usage in a sentence or relationship with meaning)
- Three types
    1. Abhidhā अभिधा (or, *primary denotation*)
    2. Lakṣaṇā लक्षणा (or, *implication*)
    3. Vyañjanā व्यञ्जना (or, *suggestion*)

# Abhidhā

- Abhidhā is the *primary* meaning
- Can be of four types
- Rūḍha (*conventional*): by convention, and may not be derived
  - ghaṭa means "pot"
- Yaugika (*etymological*): derived words
  - Using pratyaya (*suffix*): gāyaka means "sing-er", manuṣyatva means "human-ity"
  - Using samāsa (*compound word*): vidyālaya means "abode of knowledge" ("school")
- Yogarūḍha (*etymological but restricted by convention*): larger scope by etymology, but used in a restricted manner
  - paṅkaja means "lotus" (could have been anything "born in mud")
- Yaugikārūḍha (*both etymological and conventional*): by convention, and may not be derived
  - abhayā means both "one without fear" and "haritaki" (the fruit)

# Lakṣaṇā

- Lakṣaṇā is the *implied* meaning
- Can be of two types
- Nirūḍha-lakṣaṇā (*natural/unintentional implication*): original meaning is more or less dissolved and only implied meaning is used
    - kuśala means "skilled" (no longer "one who can cut kuśa grass")
- Prayojanavatī-lakṣaṇā (*intentional implication*): usage is to intend a particular meaning
    - Śuddhā-lakṣaṇā (*pure implication*): extension depends on the referrent itself
        - *inclusive implication*: "protect curd from crows"
        - *exclusive implication*: "house on Ganga"
        - *implication with partial inclusion*: "cloth is burnt"
    - Gauṇī-lakṣaṇā (*secondary implication*): extension depends on the quality of the referrent
        - *imposition of quality*: "moon-like face"
        - *total imposition*: "face is moon"

# Vyañjanā

- Vyañjanā is the *suggested* meaning
- Literal meaning is not to be understood
- “The sun has set”
  - Message to a general in battlefield: “time to attack the enemy”
  - Woman/man waiting for special friend: “time for special friend to arrive”
  - Worker in a factory to co-worker: “time to stop working for today”
  - Servant to priest: “time for evening rituals”
  - Mother to child: “stop playing and start studying”
  - Father to young daughter: “do not go far now”
  - Householder to cow-herder: “go bring back the cows”
  - Shop owner to worker: “start packing up”
  - Friend to another: “let us go out”

# Sentence

- What is a sentence?
- A *sentence* is a group of words
  - that have *mutual expectancy*, and
  - denotes a single meaning or serves a single purpose
- More mechanically, a *sentence* has one and only one *principal* verb
  - Principal verb is `tiṅanta` in Sanskrit
  - `Samāpikā Kriyā` समापिका क्रिया is one that completes
  - "The girl, sitting on the cot crying, asks her mother to bring her a toy."
- Theory of `Śābdabodha` शाब्दबोध
- Requires 3 (+ 1) factors
  - `Ākāṅkṣā` आकाङ्क्षा (*expectancy*)
  - `Yogyatā` योग्यता (*congruity*)
  - `Sannidhi` सन्निधि (*proximity*)
  - `Tātparya` तात्पर्य (*intention/purport*)

# Ākāṅkṣā

- Curiosity/expectance of a listener to hear more
- “bring”
  - what, how, where, from where
- Can be *necessary*
  - Akarmaka अकर्मक versus Sakarmaka सकर्मक verb
- Can be *mutual*
  - “door” and “close”: require each other
- Can be *one-sided*
  - “white cow”: “white” is incomplete, but “cow” is not

# Sannidhi

- Words in a single sentence should be in proximity with each other
- Words from other sentences should not be inter-twined
- “go to bring water kitchen” should better have been “Go to kitchen. Bring water.”

# Yogyatā

- Absence of obstruction in meaning
- Promotion of mutual relationship
- Lack of hindrance of valid cognition
- Ability to establish a relation
- Mutually congrous set of words
    - “sprinkle with water”
- Not congrous but cognition can still happen
    - “sprinkle with fire”
- Logically inconsistent
    - “square circle”

# Tātparya

- Intention or purport of the speaker
- Connects to `vyañjanā`
- “Bring saindhavam”
  - “Bring salt” during dinner
  - “Bring horse” in a battlefield
  - During dinner in a battlefield?
- “I didn’t beat him”
  - Purport depends on which word emphasis is put on

# Understanding a Text

- Understanding a text requires two important types of knowledge
- Language knowledge
    - Lexicon
    - Grammar
    - Pragmatics and Discourse
- Background knowledge
    - General world knowledge and common sense
    - Domain specific knowledge
    - Context
    - Culture knowledge
- Grammaticality
- Binary division: grammatical versus ungrammatical
    - “I is the …” “…seventh letter in the English alphabet”
- Scored: higher the score, the more “grammatical” it is
    - Below some low score, it stops making sense

# Tools and Resources

- Unicode is the universal standard to represent “all” *scripts*
  - *UTF-8 encoding* is the most common
  - Linux utilities gucharmap and kcharselect list the Unicode codes
- Unicode problem for Indian languages
  - A *consonant* has a halanta ending: क्, ख्, etc.
  - Unicode, however, includes the vowel marker `a' (अ) within the consonant
  - Thus, क which is grammatically 2 characters (क् + अ) is counted as 1
  - Halanta has a separate Unicode encoding as well
  - Thus, क् which is grammatically 1 character is counted as 2 (क + ◌्)
- IndicNLP AI4Bharat https://indicnlp.ai4bharat.org/home/ has a lot of resources
  - Corpora
  - ML models

# Transliteration

- Transliteration is simply changing the script
  - Not to be confused with *translation* which changes the language
- Transliteration schemes
  - Devanagari: `https://en.wikipedia.org/wiki/Devanagari_transliteration`
  - IAST (International Alphabet of Sanskrit Transliteration): `https://en.wikipedia.org/wiki/International_Alphabet_of_Sanskrit_Transliteration`
  - ITRANS (Indian Languages Transliteration): `https://en.wikipedia.org/wiki/ITRANS`
  - ISO 15919 (Transliteration of Indic Scripts): `https://en.wikipedia.org/wiki/ISO_15919`
- Transliteration tools and resources
  - Python: `https://pypi.org/project/indic-transliteration/`
  - LaTeX: `https://github.com/hrishikeshrt/devanagari-transliteration-latex`
  - Webpages: `https://aksharamukha.appspot.com/converter`