Linguistic Olympiad

boru niko	two balls	*ashi gohon*	five legs
tsuna nihon	two ropes	*ringo goko*	five cats
uma nito	two horses	*sara gomai*	five plates
kami nimai	two sheets of paper	*kaba go*	five rhinos

What will be the translation for "two plates"?

two → ni

plates → sara

nihon sara

What will be the translation for "nine cucumber"?

a) *kyuri kuhon*

b) *kyuri kuko*

c) *kyuri kuhiki*

d) *kyuri kuto*

Low Resource Languages

Overview

- Linguistic Diversity
- Why we should care?
- What can be done.

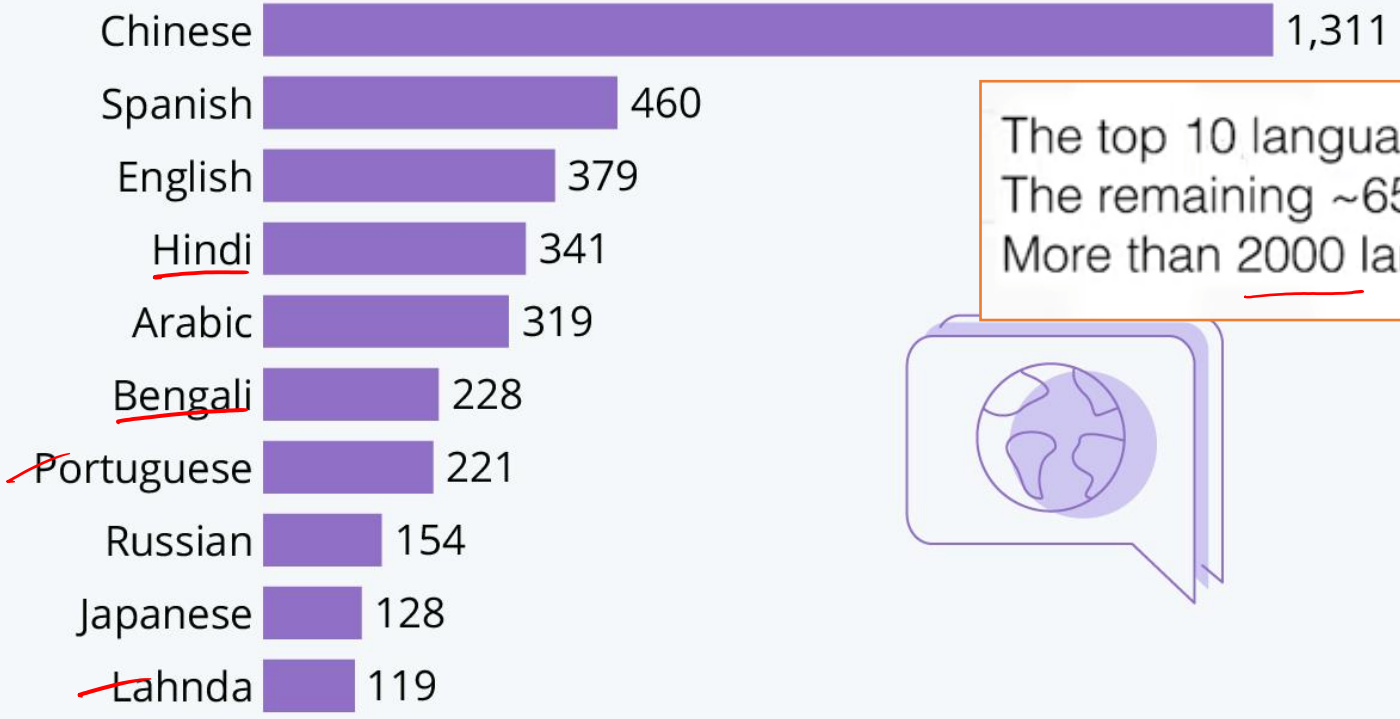
Some Stats

- 6000+ languages in the world
- 80% of the world population does not speak English
- Less than 5% of the people in the world are native English speakers.

Stanford

The World's Most Spoken Languages

Estimated number of first-language speakers worldwide in 2019 (millions)*



The top 10 languages are spoken by less than 50% of the people.
The remaining ~6500 are spoken by the rest!
More than 2000 languages are spoken by less than 1000 people.

* Each language also includes associated member languages and varieties

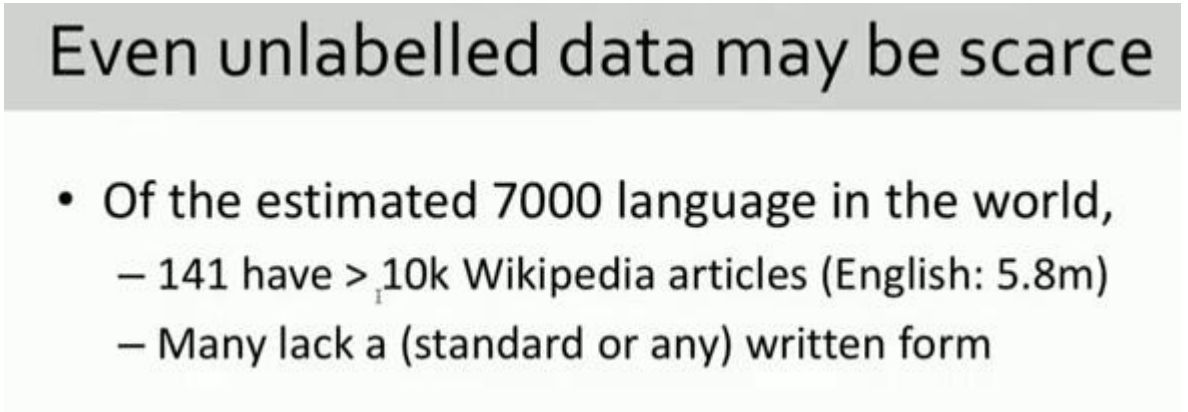
Source: Ethnologue

Low Resource Scenario

- Low Resource Language
- Low Resource Domains
- Low Resource Tasks
- Low Resource Infrastructures

Low Resource Languages

- Most languages are low -resource
  - Approximately 6,000+ languages
  - Adequate NLP resources for about 10 languages
  - Most people in the world speech a language not included in that 10
- Most domains are low -resource
  - Biomedical text
  - Legal text
  - Literary text



What are low resource languages?

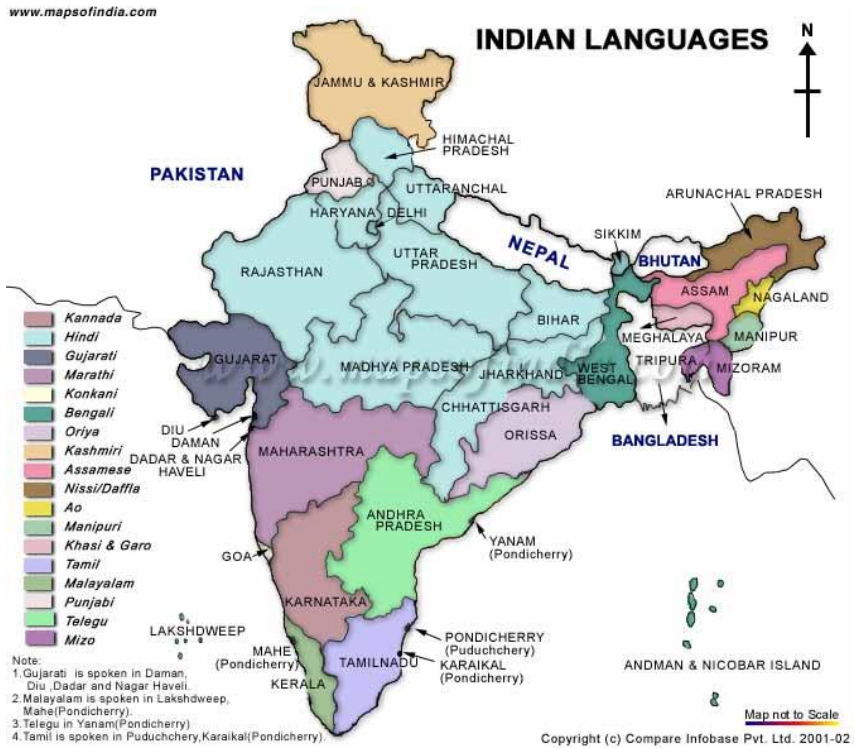
- Data
  - Raw
  - Annotated
- Tools
  - Morph. Analyzer
  - POS Tagger
  - NER
  - Chunker
  - Parser
  - MT
  - NL-Inferencing

- HUGE gap on social media (low-resource) v.s news (high-resource) text:
  - informal language and insufficient annotations

How to determine whether a language is low resource?

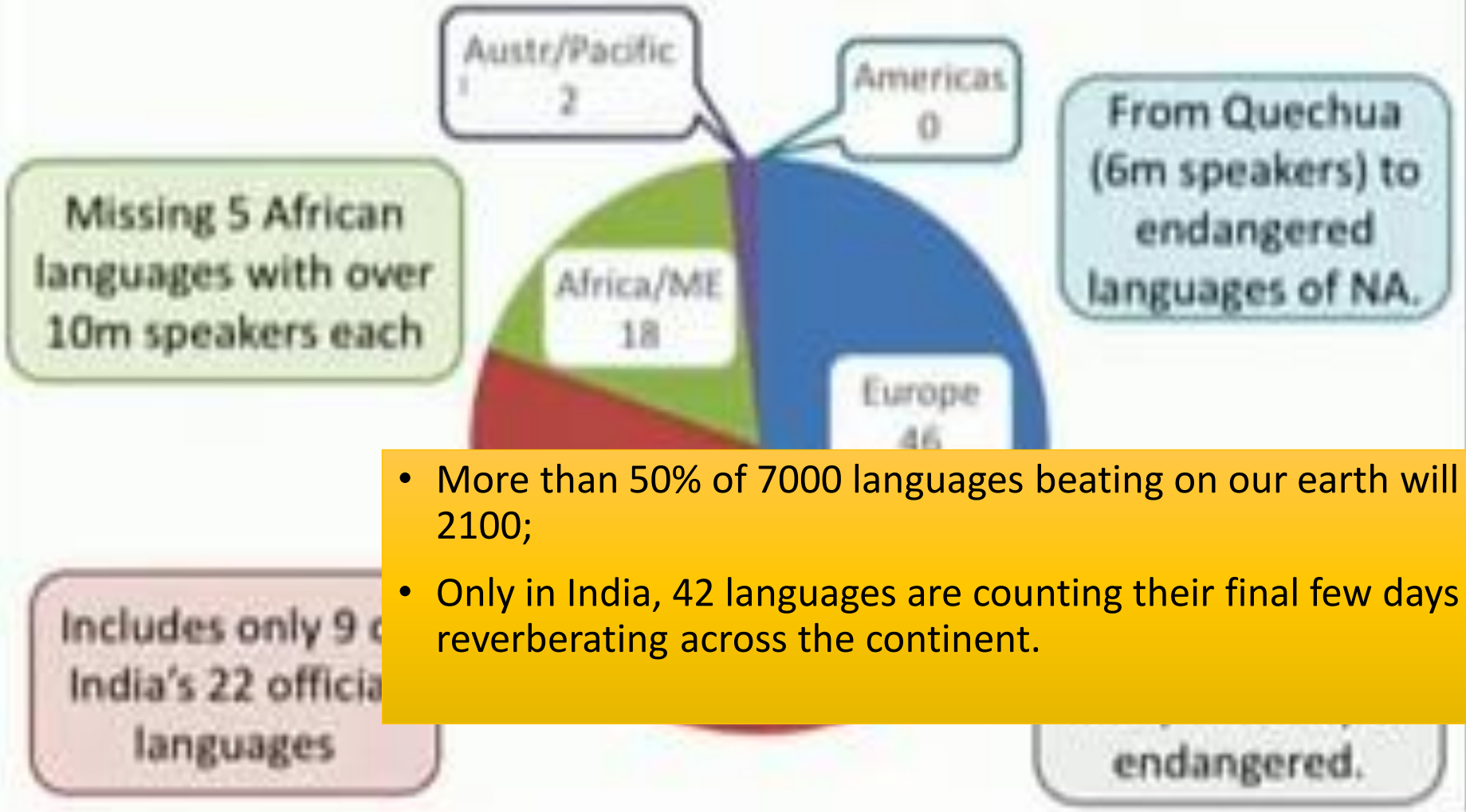
Great Linguistic Diversity

- Major streams
  - Indo European
  - Dravidian
  - Sino Tibetan
  - Austro-Asiatic
- Some languages are ranked within 20 in the world in terms of the populations speaking them
  - Hindi and Urdu: 5th (~500 million)
  - Bangla: 7th (~300 million)
  - Marathi 14th (~70 million)



Result: unequal access

- Google translate includes 103 languages.



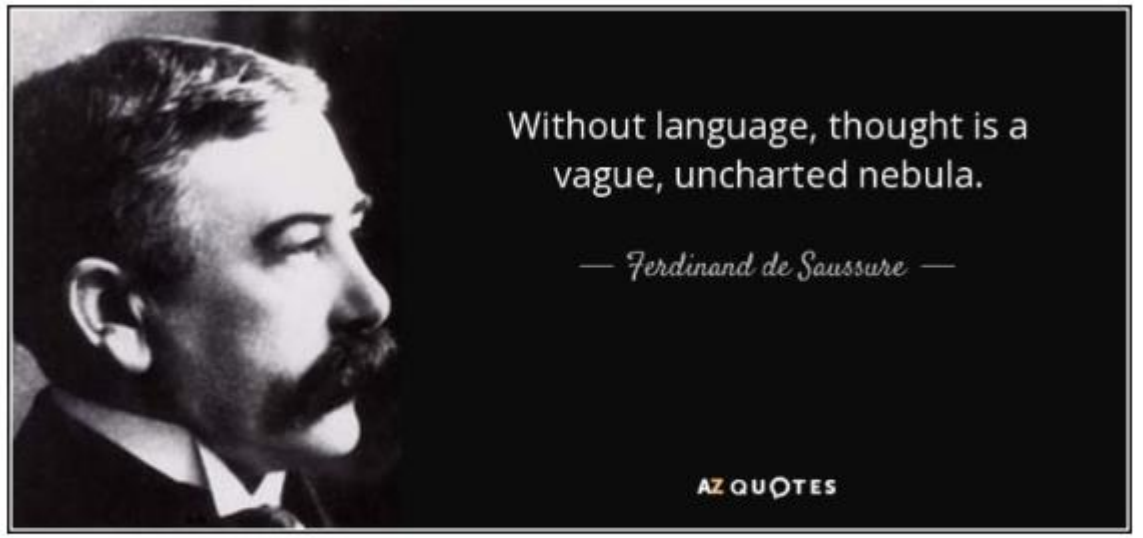
- More than 50% of 7000 languages beating on our earth will be extinct by 2100;
- Only in India, 42 languages are counting their final few days of reverberating across the continent.

So what?

So what?

When a language dies, it takes away a part of `us' with it.

How do Different Languages Influence Thought?



How do Different Languages Influence Thought?

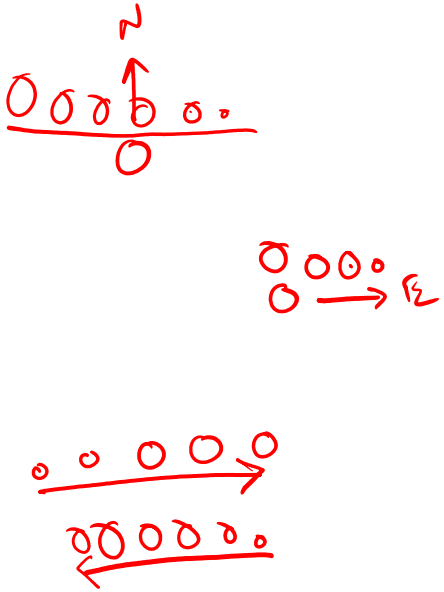
Language affects our world view and thoughts

People communicate using an array of languages, each expecting very different things from its speakers.

I saw my uncle on 42nd Street.

Language: Space and Time

Kuuk Thaayorre from Pormpuraaw



Aymara

Lera Boroditsky

https://shkrobius.livejournal.com/219313.html

The Gender

Dyirbal: Masculine, feminine, water, fire, violence, exceptional animals, vegetable and, neuter.

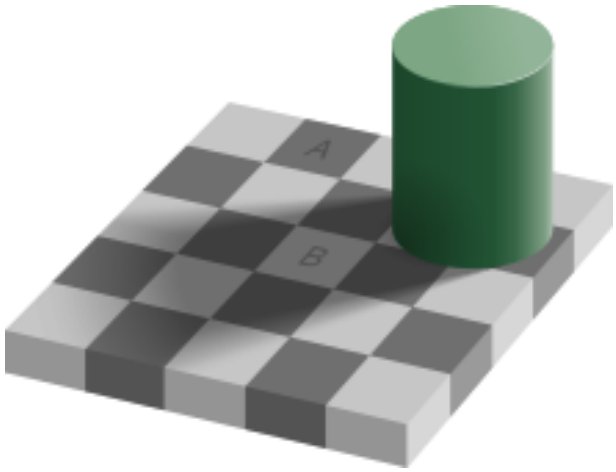
Zande: Masculine, feminine, animate, and inanimate.

Kannada- had 9 gender forms but only 3 exist at present.

What words come to your mind when you see a: Bridge ?

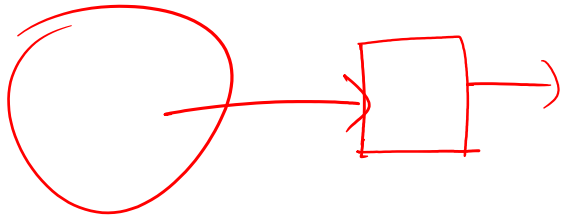
Language and Colors

Synaesthesia - a phenomenon that causes sensory crossovers, such as tasting colors or feeling sounds. Some people describe it as having “wires crossed” in their brain because it activates two or more senses when there’s only a reason for one sense to activate.



The image depicts a checkerboard with light and dark squares, partly shadowed by another object. The optical illusion is that the area labeled A appears to be a darker color than the area labeled B. However, within the context of the two-dimensional image, they are of identical brightness, i.e., they would be printed with identical mixtures of ink, or displayed on a screen with pixels of identical color.

Language and Colors



Dani from Pappuan community

Bassa spoken in Liberia

only have two color terms → Warm and Cool.

Warlpiri people living in Australia's Northern Territory don't even have any color term. For them colors are described by a rich vocabulary referring to texture, physical sensation and functional purpose.

https://nsm-approach.net/archives/6600

https://en.wikipedia.org/wiki/Aymara_language

https://en.wikipedia.org/wiki/Kuuk_Thaayorre_language#:~:text=Kuuk%20Thaayorre%20is%20similar%20to,for%20directions%20in%20Kuuk%20Thaayorre.

https://en.wikipedia.org/wiki/Languages_of_Papua_New_Guinea

Approaches

- Traditional
  - Get more data
  - Build language-specific tools with linguistic knowledge
- Unsupervised learning
  - Use machine learning techniques that do not require labeled training data

Transfer

- Exploit training data from higher-resource settings to provide supervision for low resource scenarios

Approaches

- The naivest approach to low
  - Resource scenarios is to convert them to high -resource scenarios •
    - Obtain more unannotated data
    - Annotate it
  - This has a number of obvious shortcomings
    - Raw data is often difficult to obtain.
      - Domains where only a limited amount of text exists, like law or medicine
      - Languages that do not have a significant internet presence
    - Annotation of data is expensive
      - Turkers are cheap, but unskilled and still cost money
      - Experts are expensive and slow

Rule-Based NLP

- One approach to low-resource NLP is to use models that are based on linguistic descriptions rather than being data-driven
- Given a reference grammar of sufficient quality and a lexicon, a computational linguist can build rule-based models for many things:
  - Morphological analysis
  - Parsing
  - Named entity recognition
  - Relation extraction
- However, this is also problematic
  - Not enough grammars
  - Not enough computational linguists

Linguistically Inspired ≠ Rule Based

- However, using linguistic knowledge does not mean constructing an entirely rule-based system
- One successful approach:
  - Combine linguistic knowledge and machine learning
  - Not easy with deep learning, but possible

Transfer Learning

- Learn One Place, Apply Elsewhere
- As humans, we have little problem generalizing knowledge gained in one domain to other domains
  - When we are reading legal documents, we use knowledge that we gained reading everyday English
  - When we learn Japanese, we may use knowledge that we gained speaking Korean
- This is the basic idea behind transfer learning
- It involves techniques to “transfer” knowledge gained in one domain to another

Zero-Shot and Unsupervised Learning

Training Set

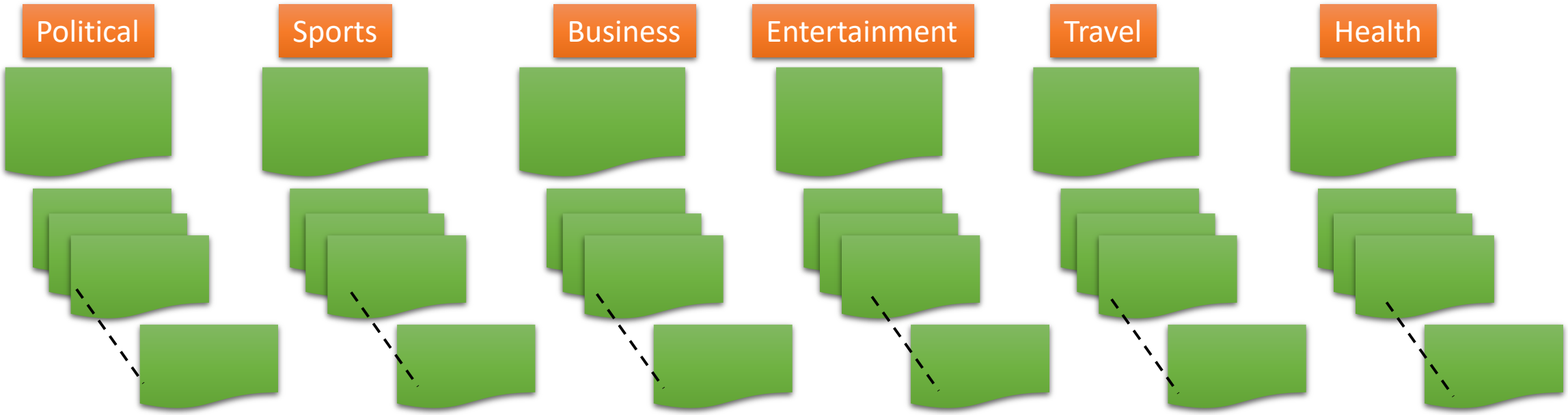
Husky Elephant Tiger

Macaw Car

Are they the same kind of animal?

Are they the same kind of animal?

News Classification



Query:

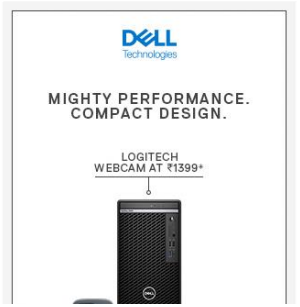
HDFC Bank Vs Kotak Mahindra Bank Vs Axis Bank: Check latest FD interest rates

By Sneha Kulkarni, ET Online • Last Updated: Mar 24, 2022, 02:47 PM IST

SHARE FONT SIZE SAVE PRINT COMMENT

Synopsis
FD interest rates have been revised by banks like Kotak Mahindra Bank, Axis Bank, and HDFC Bank. Here is a comparison of these three banks' current FD interest rates.

Bank fixed deposits are a popular alternative for risk-averse investors wishing to place their money in vehicles that offer assured returns. Bank fixed deposit interest rates differ depending on the amount invested, the period of the deposit, and senior citizens get a slightly higher rate



Support Set

By-polls Results 2022: Counting of Votes in Bengal, Bihar, Maha & Chhattisgarh Today. Full Details

West Bengal's Asansol and Ballygunge have emerged as a major prestige fight for the TMC and the BJP

NEWS18.COM • NEW DELHI
LAST UPDATED: APRIL 16, 2022, 00:12 IST
FOLLOW US ON: Facebook Twitter Instagram Telegram Google News

NEWS DESK

Counting of votes for the by-elections to one Lok Sabha seat in West Bengal's Asansol and four assembly seats in Bengal's Ballygunge, Khairagarh in Chhattisgarh, Bochahan in Bihar, and Kolhapur North in Maharashtra will begin on Saturday.

KGF 2 Tops Baahubali With Rs 134.50 Cr Earnings on Day 1; RRR Still Holds Crown for Biggest Opener

As expected, Yash and Sanjay Dutt starrer KGF: Chapter 2 is breaking records at the box office. Check out the detailed box office figures here.

NEWS18.COM
LAST UPDATED: APRIL 15, 2022, 20:54 IST
FOLLOW US ON: Facebook Twitter Instagram Telegram Google News

ENTERTAINMENT BUREAU

KGF Chapter 2, which was released on Thursday (April 14) is getting a positive response from both, the audience and the critics. As expected, the film is breaking records at the box office. As per trade analyst Taran Adarsh, KGF Chapter 2 earned Rs 134.50 crore on its opening day in India. "'KGF2' DAY 1: ₹ 134.50 CR... #KGF2 has smashed ALL RECORDS on Day 1... Grosses ₹ 134.50 cr Gross BOC [#Indibiz; ALL versions]," the trade analyst tweeted.

Bank FDs: Axis Bank, HDFC Bank, Kotak Mahindra Hike FD Interest Rates; Know Details

Recently, private sector banks, including HDFC Bank, Axis Bank and Kotak Mahindra Bank have hiked their interest rates on fixed deposits of various tenures

NEWS18.COM
LAST UPDATED: APRIL 16, 2022, 11:45 IST
FOLLOW US ON: Facebook Twitter Instagram Telegram Google News

BUSINESS DESK

Bank Fixed Deposits: Several banks, in both private and public sectors, have been hiking their fixed deposit interest rates for quite some time now. Recently, private sector banks, including HDFC Bank, Axis Bank and Kotak Mahindra Bank have hiked their interest rates on fixed deposits of various tenures and across different amounts of deposits. Fixed Deposit schemes have been the most preferred form of investment

BSCC vs KCC Dream11 Team Prediction: Check Captain, Vice-Captain, and Probable Playing XIs for MCL T20 2022 match, April 16, 1:00 PM IST

Check here BSCC vs KCC Dream11 Team Predictions and hints for the MCL T20 2022 match between Bawngkawn South Cricket Club and Kulikawn Cricket Club. Also, check the schedule of the Bawngkawn South Cricket Club vs Kulikawn Cricket Club match.

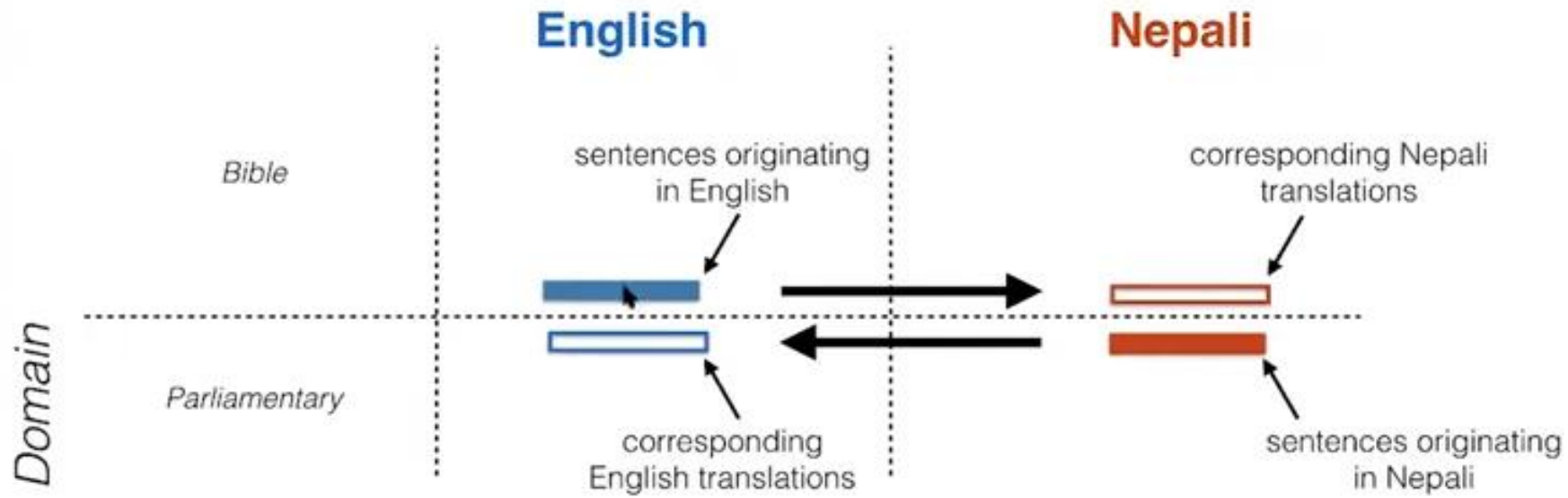
TRENDING DESK
LAST UPDATED: APRIL 16, 2022, 12:10 IST
FOLLOW US ON: Facebook Twitter Instagram Telegram Google News

BSCC vs KCC Dream11 Team Prediction and Suggestions for today's MCL T20 2022 match between Bawngkawn South Cricket Club and Kulikawn Cricket Club:

Bawngkawn South Cricket Club and Kulikawn Cricket Club will square off against each other in the upcoming game of the MCL T20 2022 on Saturday, April 16. The two teams will play at the Suaka Cricket Ground in Mizoram. Both Bawngkawn South Cricket Club and Kulikawn Cricket Club need to make some amends to ensure a better ride in the league.

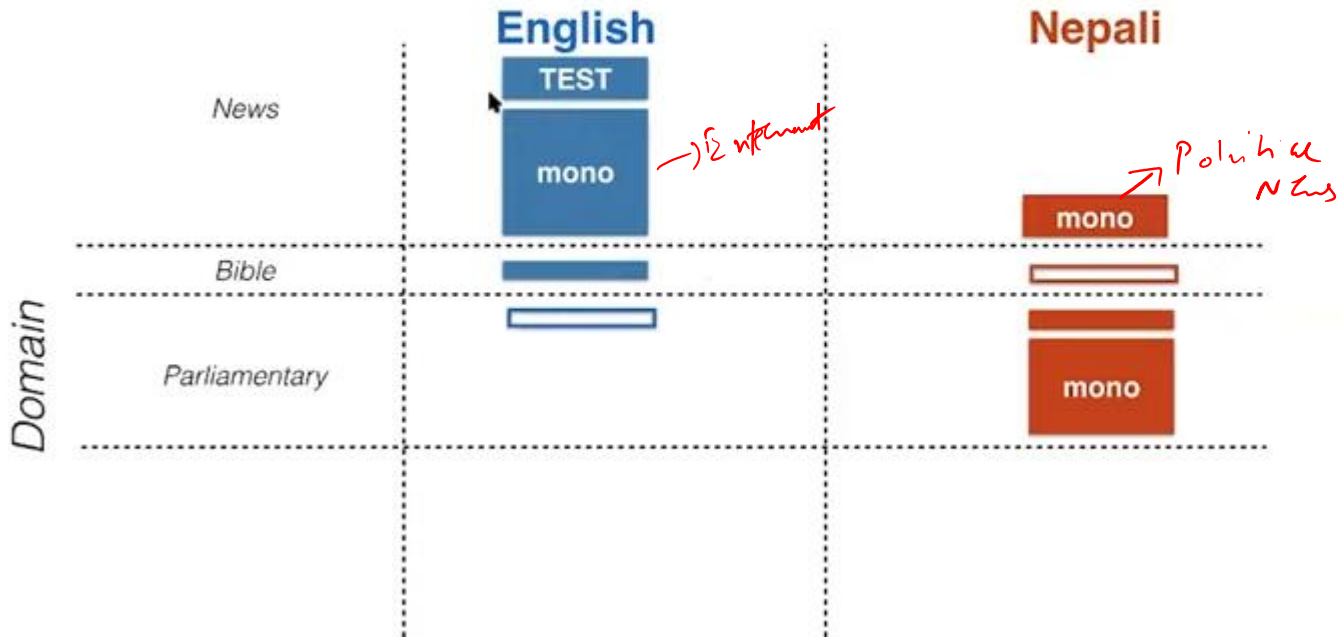
Bawngkawn South Cricket Club are reeling at the bottom of the points table. They have lost all three games and are yet to open their account in the competition. Bawngkawn's most recent loss in the competition came against CVCC by 48 runs. The batters let the team down as they scored only 99 runs while chasing 148.

Machine Translation in Practice

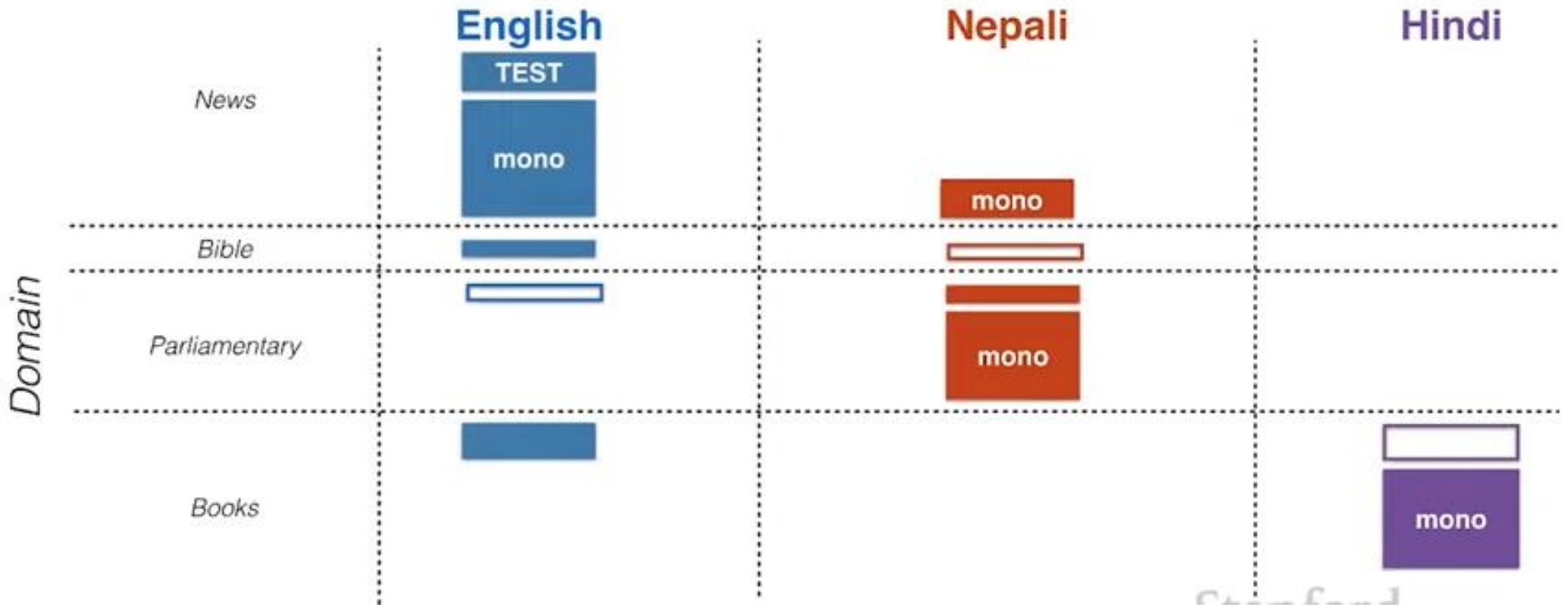


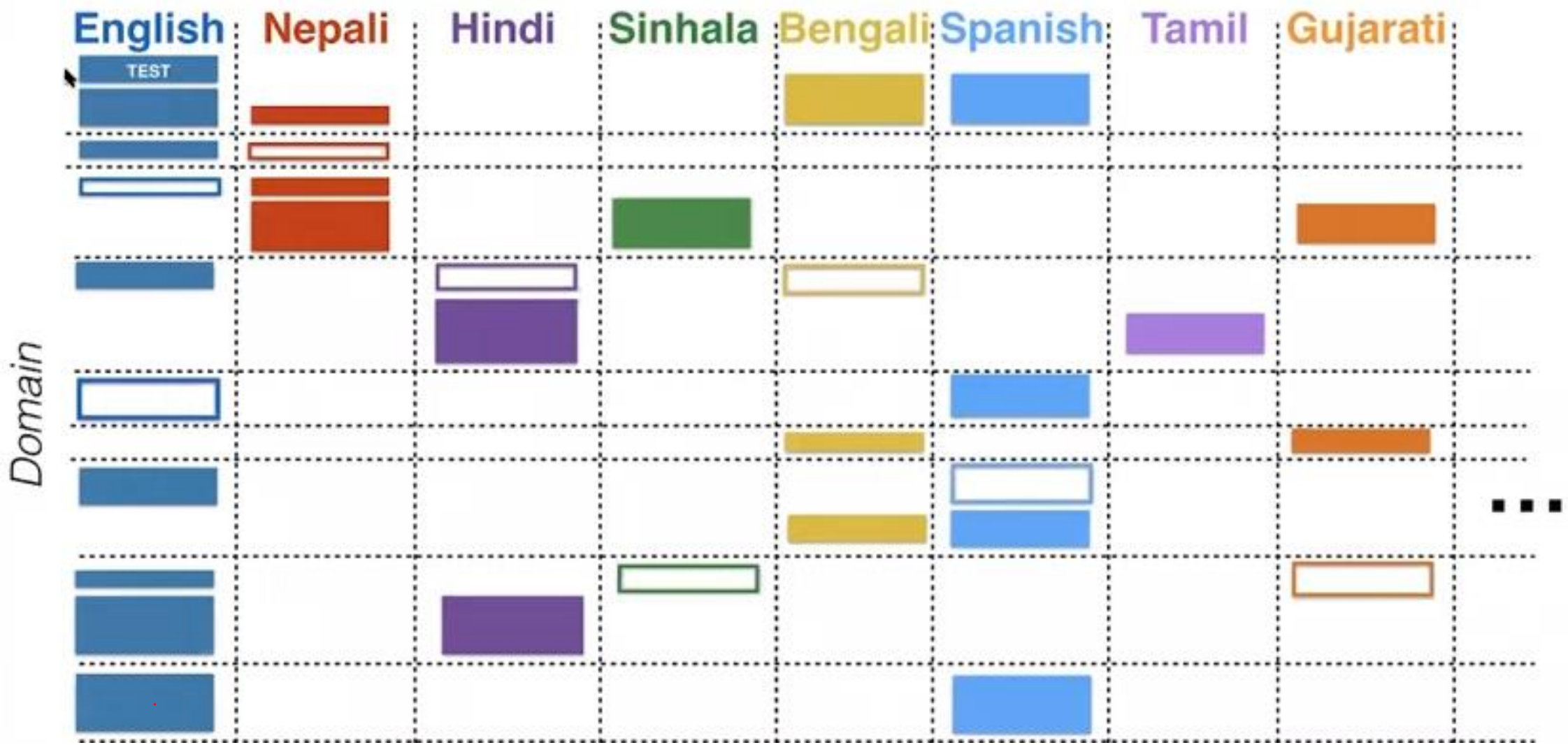
Let's represent (human) translations with empty rectangles.

Machine Translation in Practice



Machine Translation in Practice





Mondrian Like Language Setting