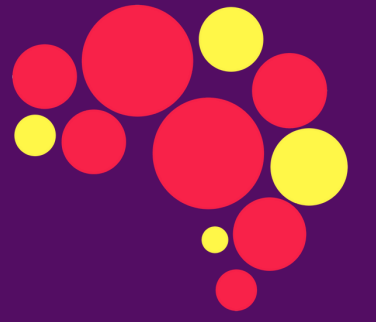


# NLP for Low Resource Languages





**Ms. Gloriana Monko**

**Department CSE-UDOM**

**Liaison Manager - AI4D Anglophone African Lab**

**Co-founder TelesoftAI, WS2**

**PhD Student: Shibaura Institute of Technology-Tokyo**



# The Future of Communication

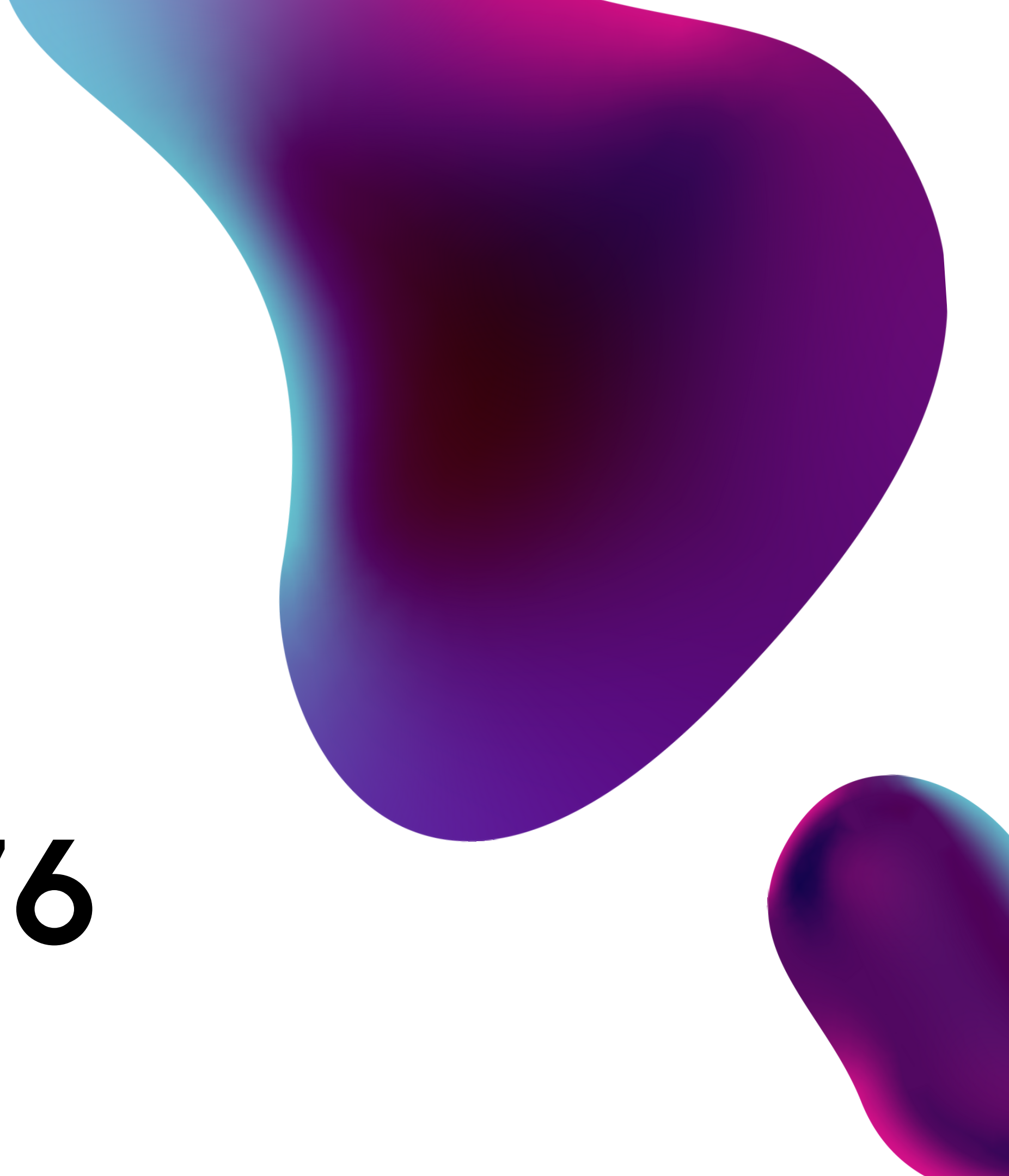
What's in store for us?

**Communication is central in how  
we live.**

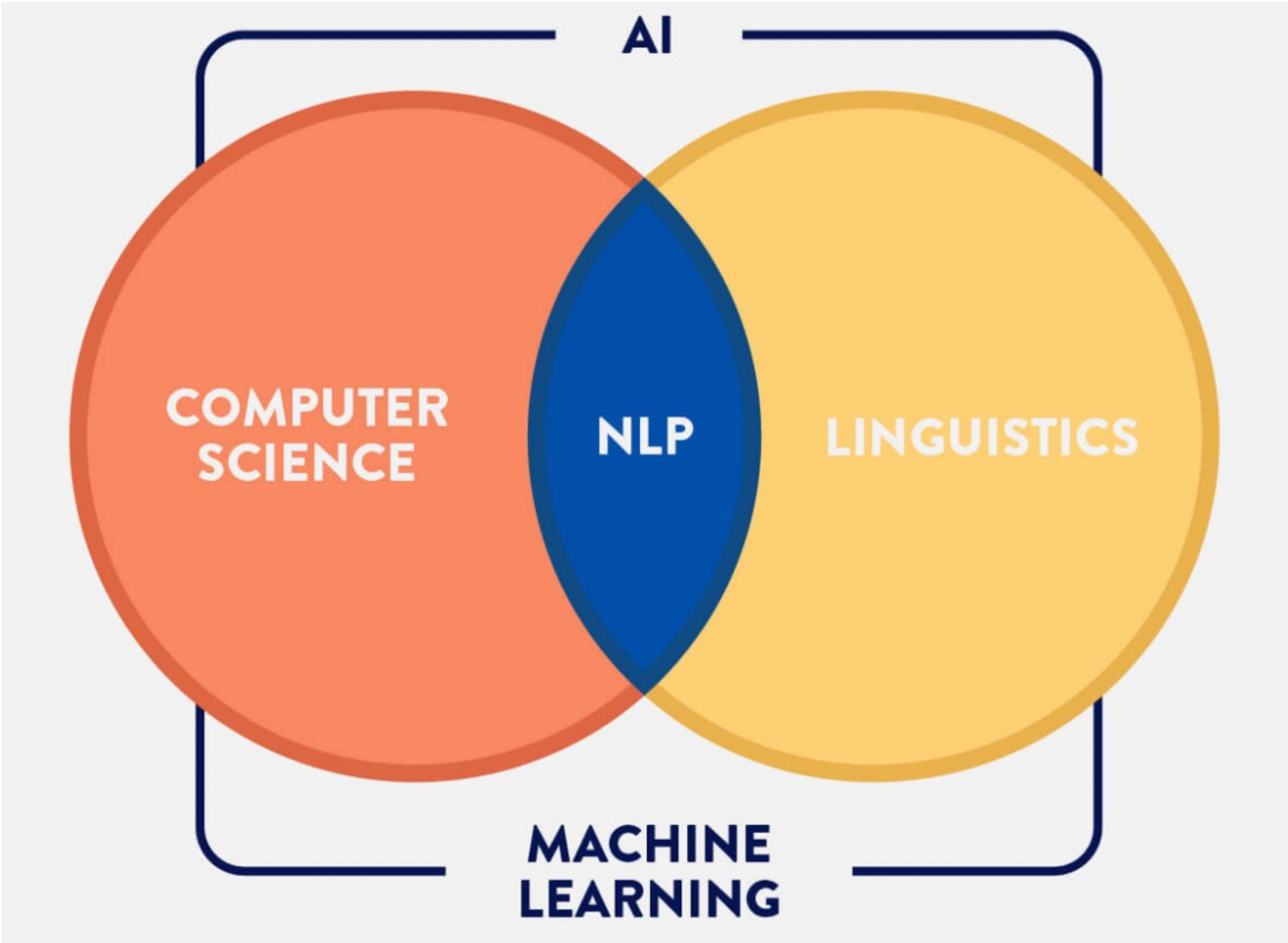
**Go to**

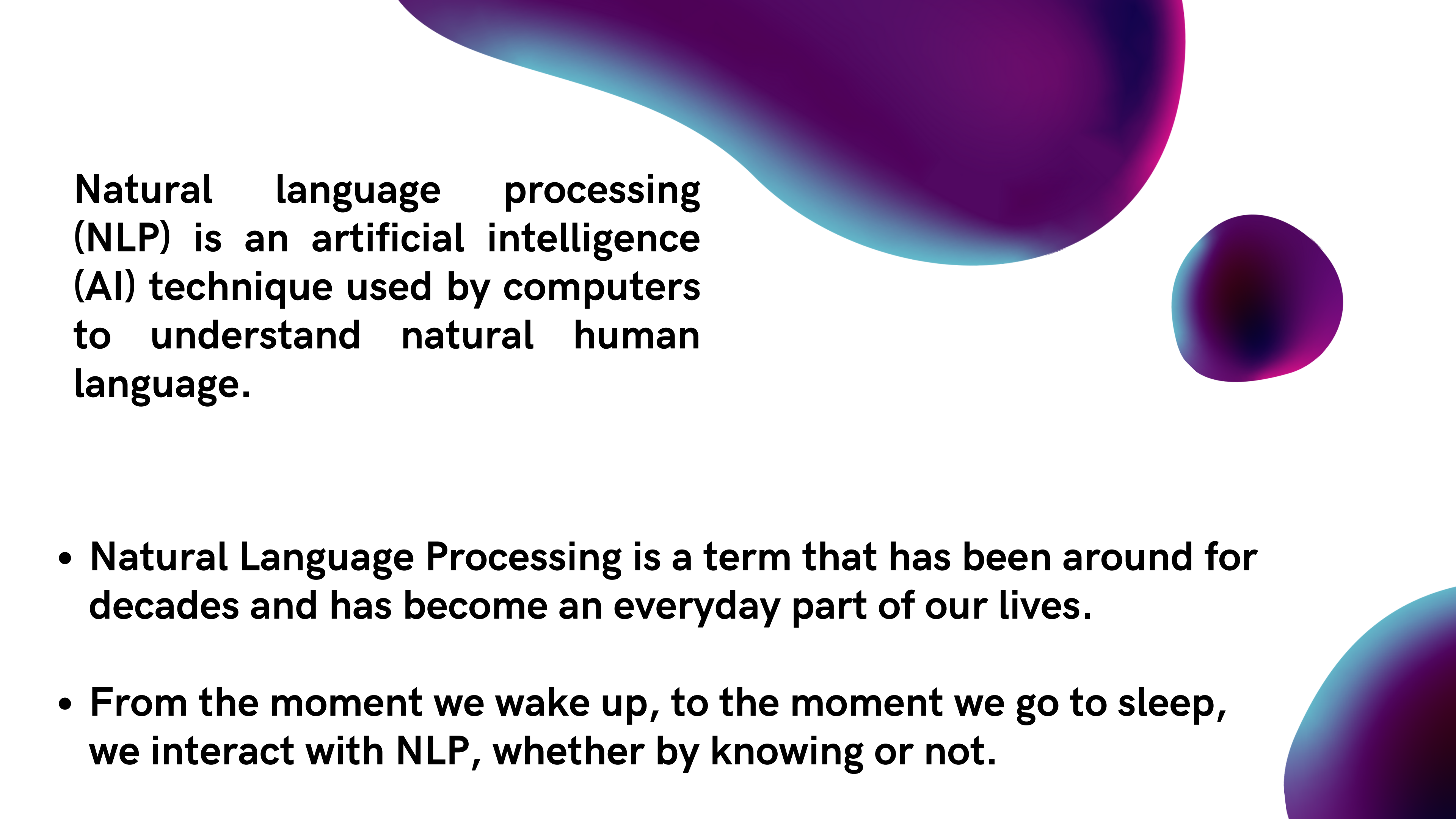
**menti.com**

**code: 43967276**









**Natural language processing (NLP) is an artificial intelligence (AI) technique used by computers to understand natural human language.**

- Natural Language Processing is a term that has been around for decades and has become an everyday part of our lives.**
- From the moment we wake up, to the moment we go to sleep, we interact with NLP, whether by knowing or not.**

# NLP Presence

Machine Translation

Information Retrieval

Question Answering



Dialogue Systems

RoboChats

Summarization



Information Extraction

Speech recognition

Sentiment Analysis



# Proliferation of Chat Bots



**Deliver faster chat support**



**Set up easily with bot templates**



**Cut down on support costs**

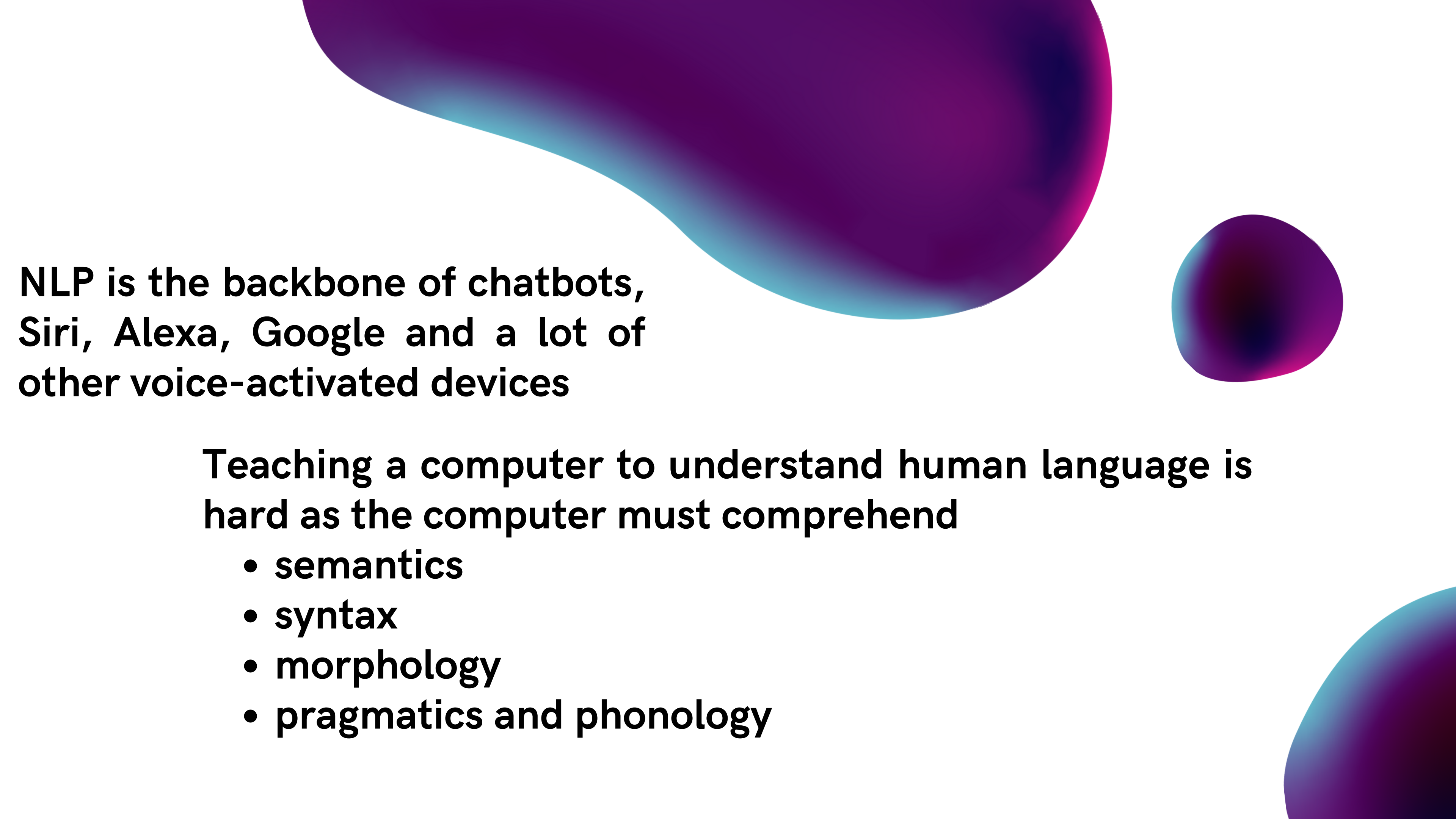


**Offer proactive customer service**



**Optimize your chatbot's performance**





**NLP is the backbone of chatbots, Siri, Alexa, Google and a lot of other voice-activated devices**

**Teaching a computer to understand human language is hard as the computer must comprehend**

- **semantics**
- **syntax**
- **morphology**
- **pragmatics and phonology**

# Why NLP is hard?

- **Ambiguity at many levels:**
  - bank (finance or river?)
  - chair (noun or verb?)
  - Syntactic structure: I saw a man with a telescope
  - Quantifier scope: Every child loves some movie
- **Linguistic diversity**

# Main Focus

**01**

**What is low-resource NLP?**

**02**

**Why low-resource NLP is hard?**

**03**

**Why care about low-resource languages?**

**04**

**Approaches to low-resource NLP**





**In NLP-languages are often referred as low resource or high resource**

**High resource languages**

Many data resources exist, making possible the development of machine-learning based systems for these languages. Eg. English

**Low resource languages**

The languages with none or very few resources available. Eg. Swahili



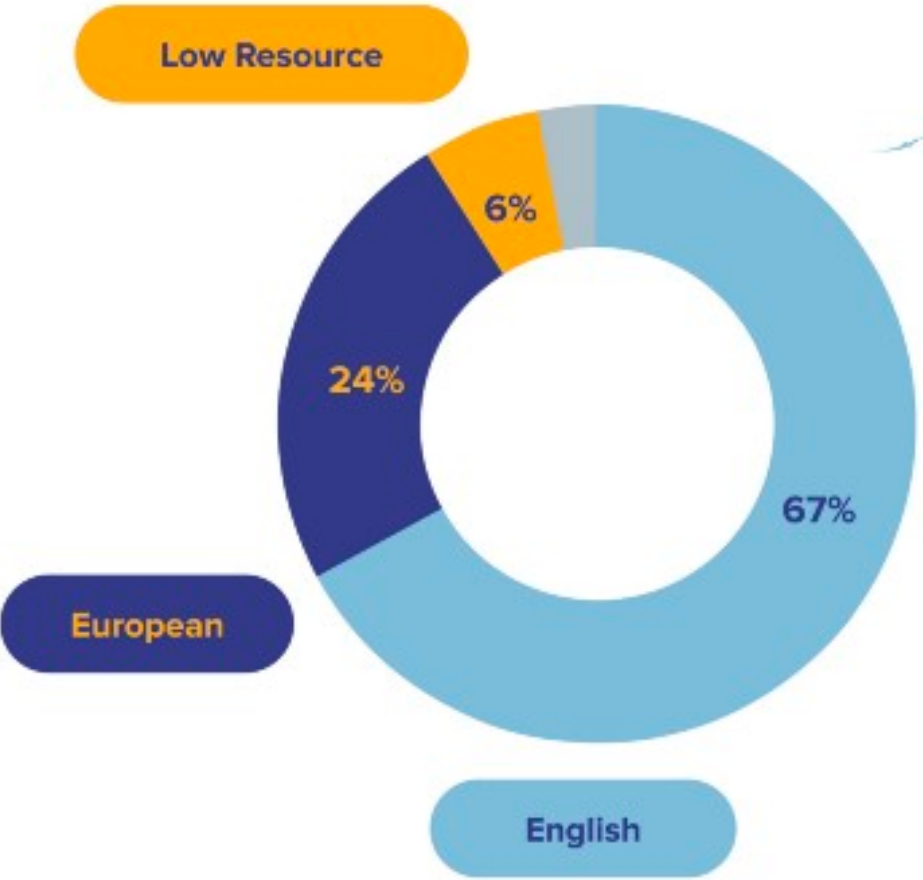


# Why low-resource NLP is hard?

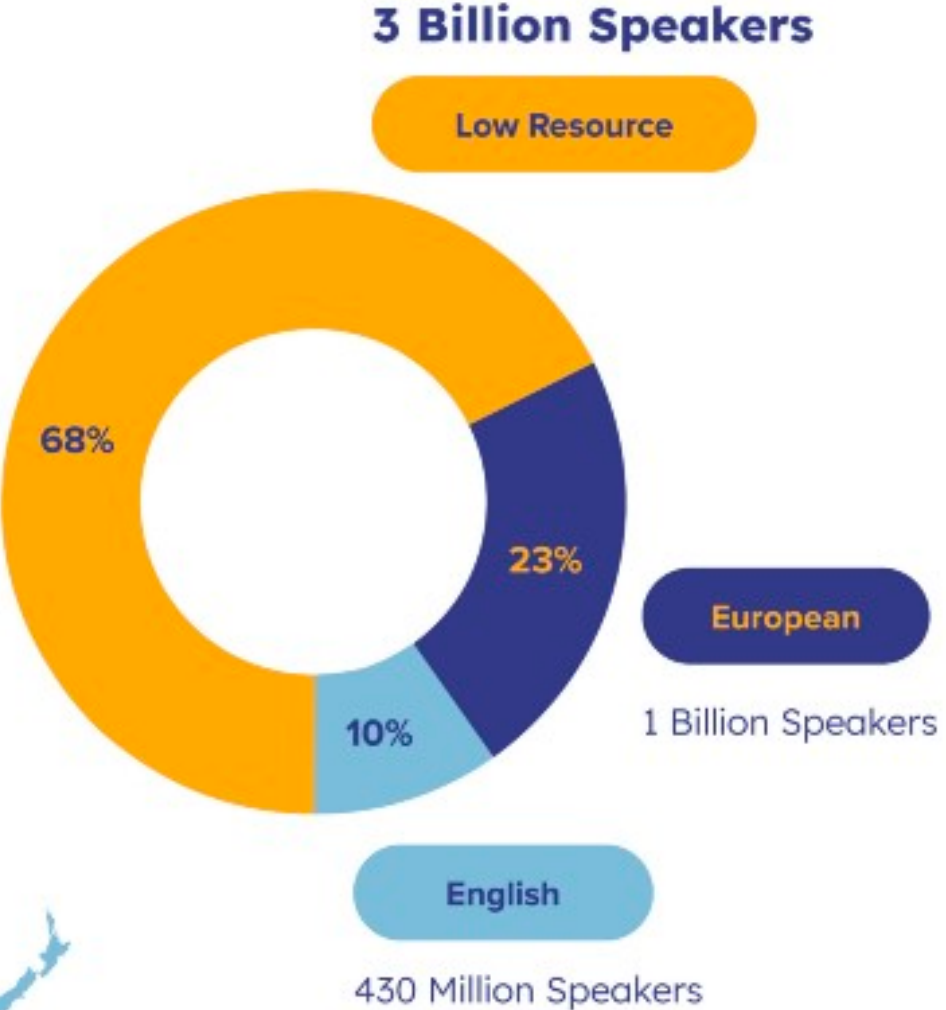
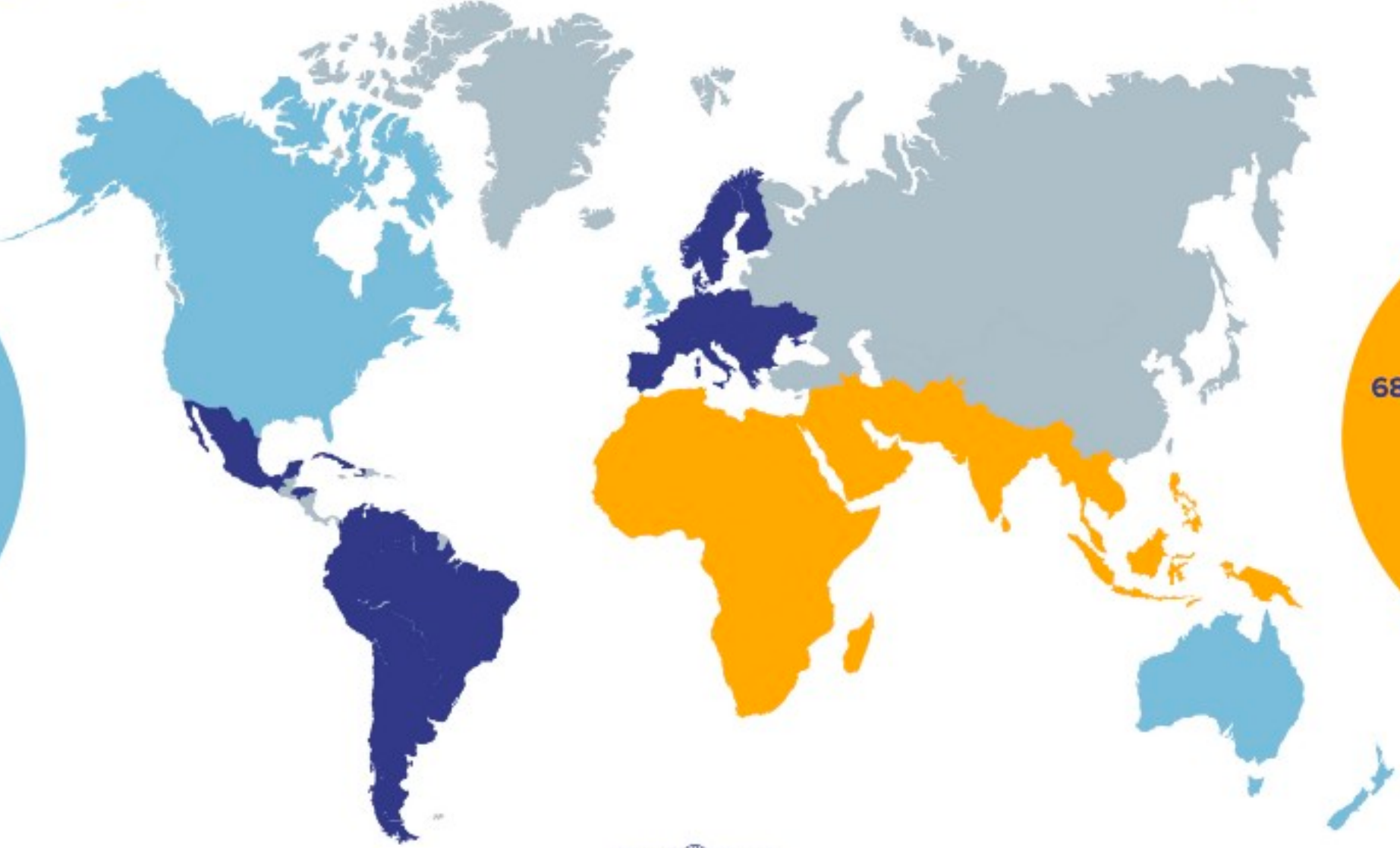
It is even harder for so-called low-resource languages where the annotated resources are very limited.

There are approximately 7,000 languages in the world, but of these only a small fraction (20 languages) are considered high-resource languages.

# NLP Solutions by Language



# Population Size of Languages



*Published in NeuralSpace-2022*

# Challenges of Low-Resource NLP

- Software products containing NLP features are estimated to globally generate USD 48 billion by 2026 (NeuralSpace, 2022).
- Current NLP solutions majorly focus on English, Spanish or German although there are about 3 billion low-resource language speakers (mainly in Asia and Africa)

**Lack of annotated datasets**

**Lack of unlabelled datasets:**

**Supporting multiple dialects of a language:**

# Why care about low-resource languages?

The scarcity of parallel data is a major obstacle for training high-quality machine translation systems for low-resource languages.

Low-resource languages are in dire need of tools and resources to overcome the resource barrier such that advances in NLP can deliver more widespread benefits.



# Why care about low-resource languages?

Fortunately, some low-resource languages are linguistically related or similar to high-resource languages; these related languages may share many lexical or syntactic structures.

# Leveraging High-resource Languages to Improve Low-resource Translation

## Parallel data

It is beneficial to mix the limited parallel data pairs of low-resource languages with high-resource language data. Lakew et al. (2019)

## Multilingual Learning

Where single model is trained on multiple languages.

## Data Augmentation

The strategy that automatically creates new data without collecting it explicitly.

# Transfer learning

Provides an important opportunity for low-resource NLP, whereby annotation is transferred from a source resource-rich language to a target resource poor-language.

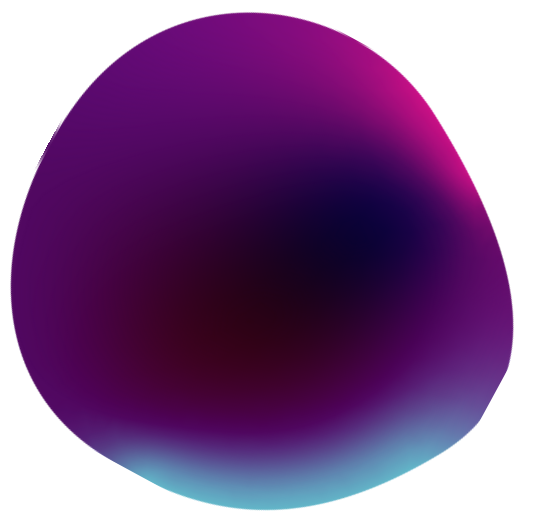
The use of Pre-trained models such as BERT-Hugging Face





# Auto Translation

No longer will you have to learn a foreign language to communicate with others.





"Computers are incredibly fast, accurate and stupid;  
humans are incredibly slow, inaccurate and brilliant;  
together they are powerful beyond imagination."

Albert Einstein

# Gloriana Monko



**gloriana.monko@ai4dlab.or.tz**



**gmonko24@gmail.com**



**@GlorianaMonko**



**Gloriana Monko**

**AI4D website: [www.ai4dlab.or.tz](http://www.ai4dlab.or.tz)**

**TelesoftAI : <https://telesoftai.com/#>**

A word cloud featuring the phrase 'thank you' in various languages and scripts. The central and largest text is 'thankyou' in red. Other prominent words include 'danke' (German), '感謝' (Japanese), 'ngiyabonga' (Ndebele), 'tesekkür ederim' (Turkish), 'gracias' (Spanish), 'mochchakkeram' (Bhojpuri), 'go raibh maith agat' (Irish Gaelic), 'sukriya' (Hindi), 'kop khun krap' (Lao), 'arigatō' (Japanese), 'tak' (Tamil), 'dakujem' (Slovak), 'merci' (French), 'obrigado' (Portuguese), 'dziękuję' (Polish), 'bedankt' (Dutch), 'спасибо' (Russian), 'rahmat' (Indonesian), '감사합니다' (Korean), 'merci' (Italian), 'misaotra' (Malagasy), 'matondo' (Malawi), 'paldies' (Latvian), 'grazzi' (Italian), 'tapadh leat' (Irish Gaelic), 'hvala' (Slovene), 'mauruuru' (Māori), 'kösönöm' (Hungarian), 'dhanyavad' (Thai), 'nannri' (Hawaiian), 'nandri' (Hawaiian), 'kiitos' (Finnish), 'dankie' (Afrikaans), 'faafetai lava' (Samoan), 'vaka' (Fijian), 'spasi' (Fijian), 'blagodaram' (Sanskrit), 'kia ora' (Māori), 'barka' (Arabic), 'welalin' (Hausa), 'tack' (Swedish), 'misaotra' (Malagasy), 'matondo' (Malawi), 'paldies' (Latvian), 'grazzi' (Italian), 'mahalo' (Hawaiian), 'tapadh leat' (Irish Gaelic), 'xвала' (Ukrainian), 'asante' (Kisumu), 'manana' (Hawaiian), 'obrigada' (Portuguese), 'chokrane' (Hindi), 'muraqaze' (Pashto), 'tenki' (Hindi), 'djiere dieuf' (Sango), 'tau' (Tamil), 'mamnun' (Urdu), 'sulpay' (Tibetan), 'chnorakaloutioun' (Khmer), 'gracias ago' (Portuguese), 'gracies' (Catalan), 'sagolun' (Korean), 'didi madloba' (Hindi), 'kam sah hamnida' (Burmese), 'najs tuke' (Slovak), 'arigatō' (Japanese), 'tanemirt' (Slovak), 'rahmet' (Turkish), 'diolch' (Welsh), 'dhanyavadagalu' (Tamil), 'shukriya' (Urdu), 'merce' (Catalan), 'merci' (French), 'xiexie' (Chinese), and 'ευχαριστώ' (Greek).