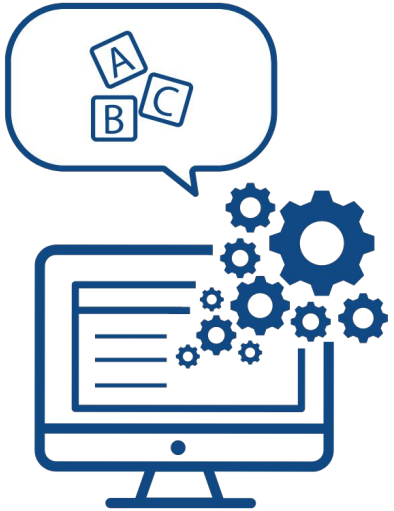
Leveraging the **NLTK Library** for **Translation**

*A Case Study for **Dyula- French Translation***

Praelexis

NLP

Natural Language Processing



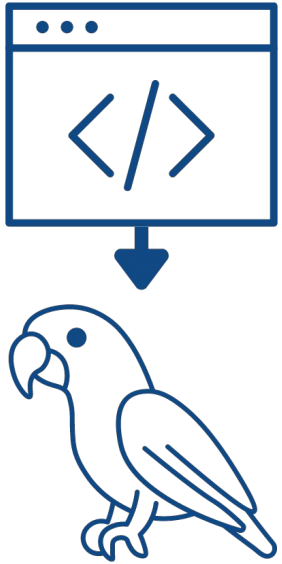
(1)Natural language processing (NLP) is a subfield of computer science and artificial intelligence (AI) that uses machine learning to enable computers to understand and communicate with human language.

NLP

Natural Language Processing

(2) The art of making computers understand human language.

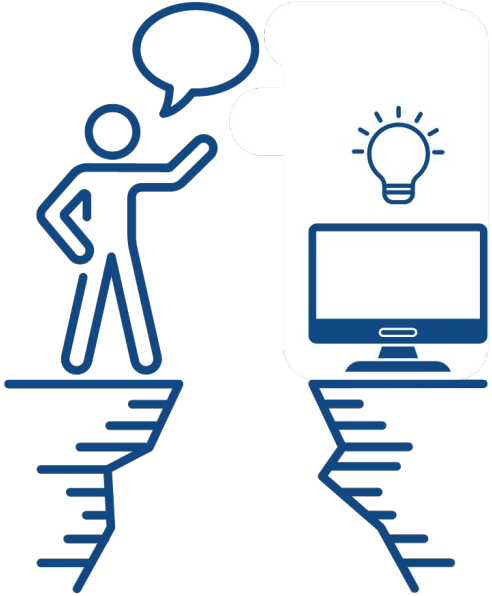
It is like teaching your pet parrot to speak, but with code!



NLP

Natural Language Processing

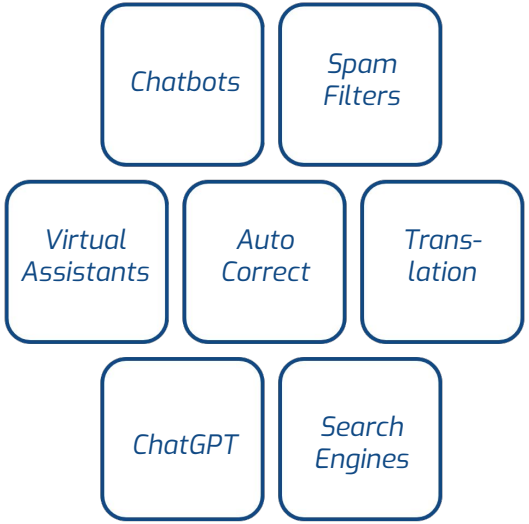
(3) NLP is the bridging the gap between human communication and computer understanding



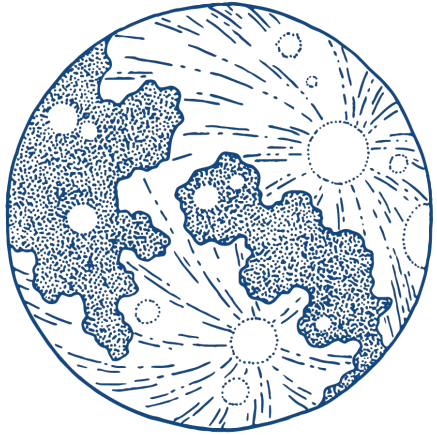
NLP

Natural Language Processing

(4) NLP is used behind many of our tools we use today

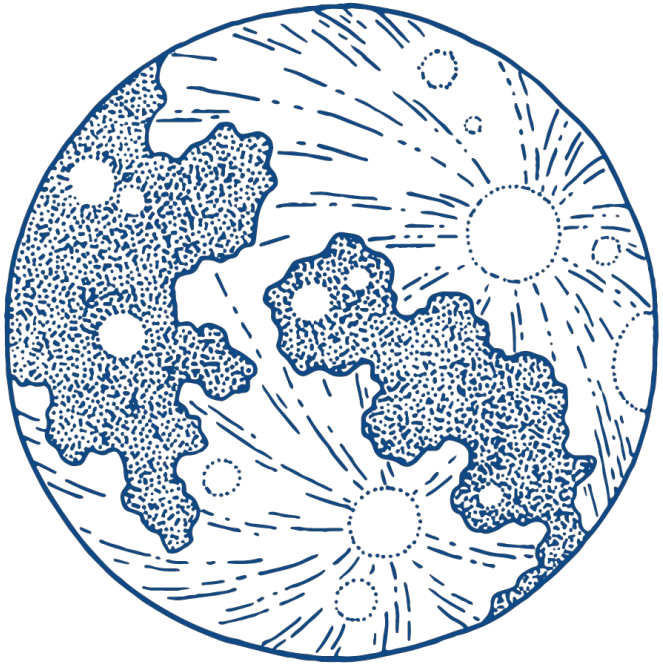


Did an incorrect translation lead to the search for life on Mars?



Why is translation so difficult?

In 1877, an Italian astronomer named Giovanni Schiaparelli noted he had seen what appeared to be ***canali***, on **Mars**. This term, discovered some years later in his writing, was interpreted to mean canals and sent budding scientists scrambling to identify the life on Mars that could have created such ***canals***. Unfortunately, the Italian word canali is just a general term to describe trenches, which can be part of the natural terrain and not necessarily man-made. The idea of life on Mars, however, has long outlived the legend of this mistranslation.



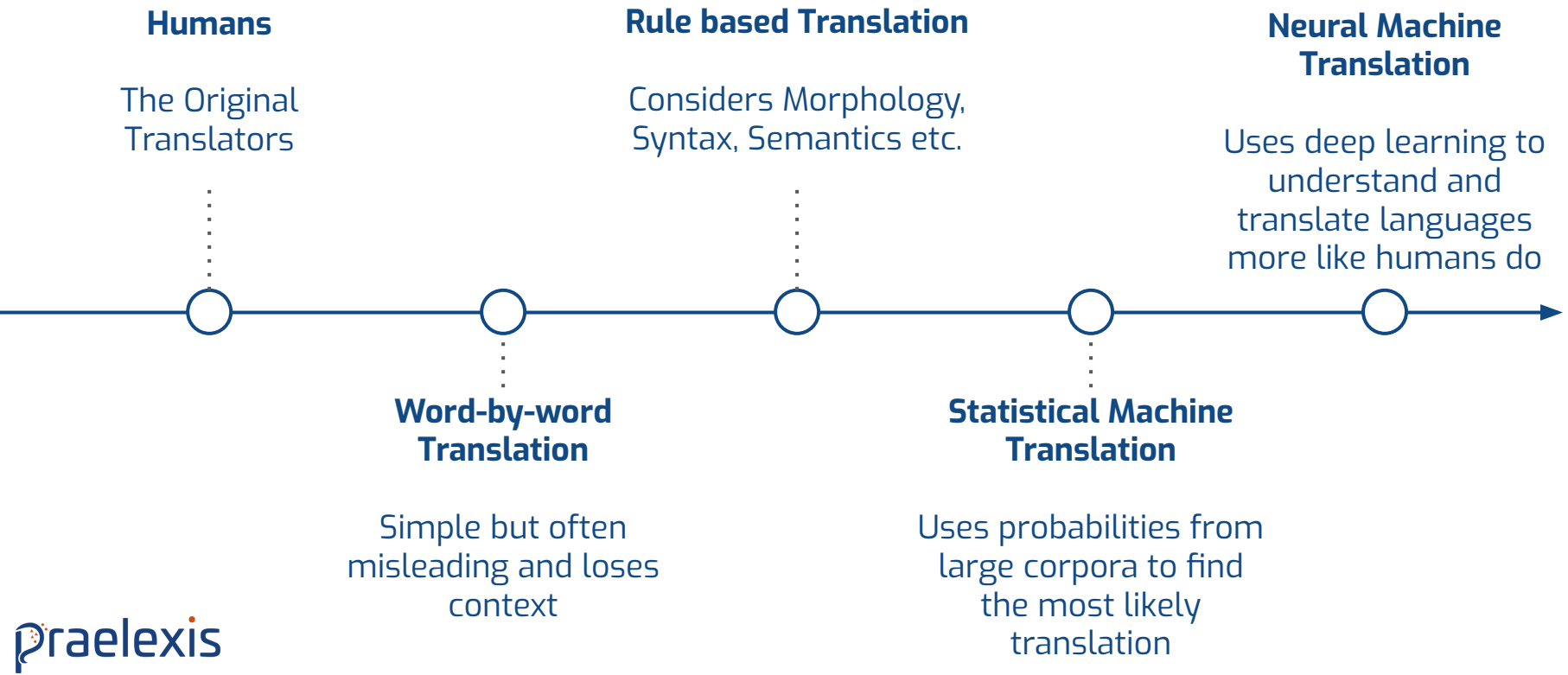
Why is translation so difficult?

"Languages are like fingerprints—each one unique and full of quirks."

Why is translation so difficult?

1. Languages have **colorful** descriptions and idioms, with the meaning not always literal
2. **Context** matters: Words can have different meanings in different situations
3. There can also be **cultural differences**
4. Translation is **not just** about **words**—it's about capturing the culture. A 'simple' greeting can mean have different meaning in different contexts.
5. **Lack** of **Data**
6. It is difficult to create translation datasets for **rare languages**

How is translation normally done?



Praelexis

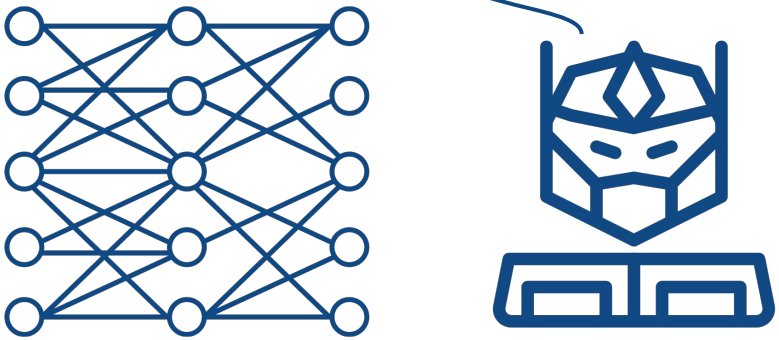
What libraries are available?

Library	Main Features	Best for	Pros	Cons
NLTK	Tokenization, stemming, POS tagging, translation	Educational use, in-depth NLP exploration	Comprehensive, Flexible, Built-in corpora	Older design, Slow
SpaCy	Entity recognition, POS tagging	Large-scale processing, speed-critical tasks, Production applications	State-of-the-art models, Fast and efficient, User-friendly API,	Limited flexibility,
TextBlob	Sentiment analysis, basic translation	Quick and simple tasks, beginners	Simple API	Basic functionality,
Transformers	Cutting-edge models (BERT, GPT, T5), deep learning	High-quality translation, state-of-the-art NLP	State-of-the-art models, Pretrained models	Resource-intensive, Complex
Deep Translator	Wrapper for Google, Microsoft, Yandex, etc. translation APIs	API-driven translation tasks (wrapper for various APIs)	Multiple translation provider	Limited to third party APIs

What libraries are available?

Library	Main Features	Best for	Pros	Cons
NLTK	Tokenization, stemming, POS tagging, translation	Educational use, in-depth NLP exploration	Comprehensive, Flexible, Built-in corpora	Older design, Slow
SpaCy	Entity recognition, POS tagging	Large-scale processing, speed-critical tasks, Production applications	State-of-the-art models, Fast and efficient, User-friendly API,	Limited flexibility,
TextBlob	Sentiment analysis, basic translation	Quick and simple tasks, beginners	Simple API	Basic functionality,
Transformers	Cutting-edge models (BERT, GPT, T5), deep learning	High-quality translation, state-of-the-art NLP	State-of-the-art models, Pretrained models	Resource-intensive, Complex
Deep Translator	Wrapper for Google, Microsoft, Yandex, etc. translation APIs	API-driven translation tasks (wrapper for various APIs)	Multiple translation provider	Limited to third party APIs

When should I consider alternatives?



Representation of an actual transformer

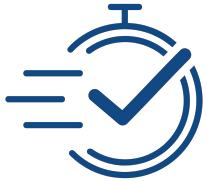
Representation of a transformer model

Transformers have become the go to for translation

Great for **complex translations** and **high accuracy**, especially when **large amounts** of **data** are available.

But... sometimes you need something ***lighter***, ***faster***, *or* ***simpler...***

When to use other libraries

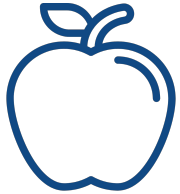


When **speed** matters

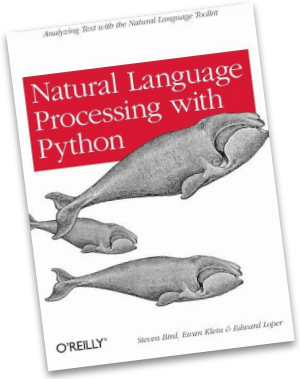
When you have **limited data**

When you have **resource constraints**

When you want a **fast implementation** (easy set-up)



When you have a **m1 chip** on your **mac** and struggle with your environment



NLTK toolkit for **NLP tasks**

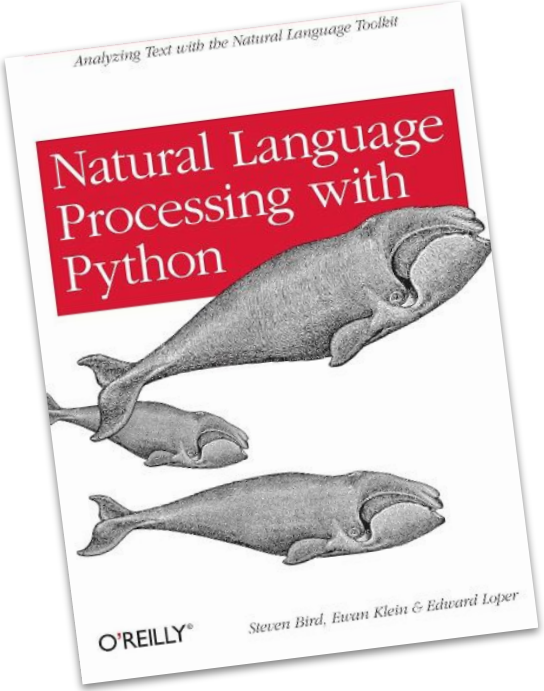
Praelexis

NLTK

Natural Language Toolkit

(1) NLTK (Natural Language Toolkit) is a powerful Python library for working with human language

Natural Language Processing with Python,
by Steven Bird, Ewan Klein, and Edward Loper
(also available online at https://www.nltk.org/book/)



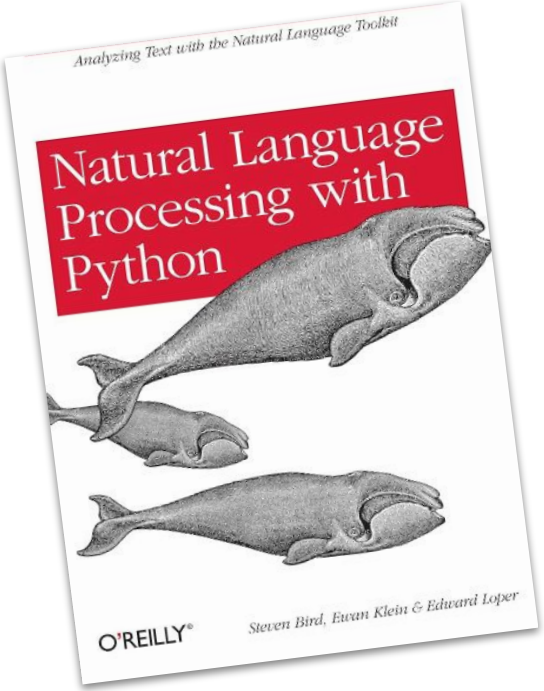
Praelexis

NLTK

Natural Language Toolkit

(2) It was developed in 2001 as part of a research project at the university of Pennsylvania

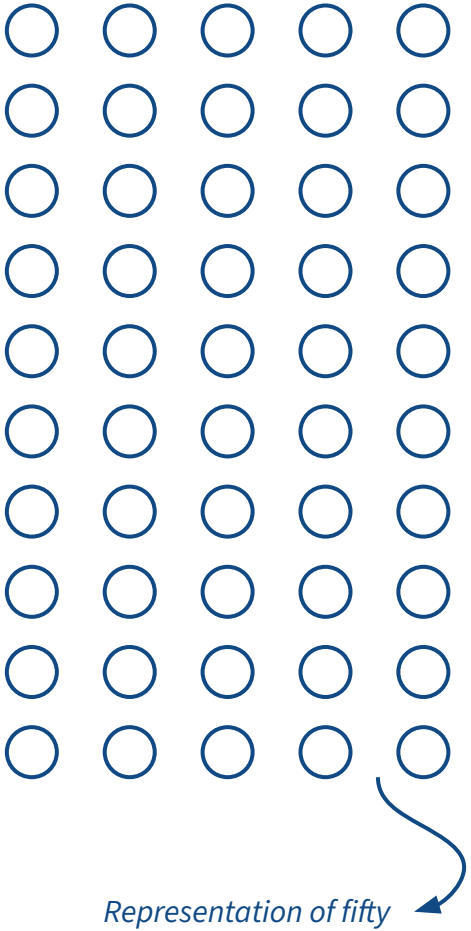
(3) Still widely used as a teaching and research tool



NLTK Library

Natural Language Toolkit Library

(1) The NLTK library has
over **50 corpora** and **lexical resources**



Representation of fifty

Praelexis

Corpora

a text corpus is a large body of text

---

(1) The NLTK library includes Corpora in different languages and styles

*(webchat, books, movie reviews, shakespeare texts, and most known **WordNet**)*

Lexical Resources

A collection of words and/or phrases along with associated information such as parts of speech and sense definition

---

(1) Wordlist Corpora useful for solving Wordle or for spell checkers
(2) Pronouncing Dictionaries (CMU Pronouncing Dictionary for US English)
(3) **Wordnet:** Synonyms, Semantic Similarity etc.

***NLTK is modular**—each module focuses on different tasks.*

Linguistic Datasets

nltk.corpus

Linguistic datasts

Tokenization

nltk.tokenize

Split text into smaller components

Example:
[My, name, is, Alta]

Stemming & Lemmatization

nltk.stem

Reduce words to their root forms

Example:
Running → Run
Ran → Run

POS tagging

nltk.pos_tag

Labeling words in a Sentence

Example:
[('My', 'PRP$'), ('name', 'NN'), ('is', 'VBZ'), ('Alta', 'NNP')]

Translation

nltk.translate

Models and Metrics

NLTK toolkit for translation

nltk.corpus

nltk.translate

Comparative Word Lists

Swadesh Wordlists (24 languages)

Spanish	French
eau	agua
hombre	homme

Alignments

Aligns words in parallel sentence pairs

Example:

src : ['mun', 'fɛn', 'dɔ']
trg: ['quoi', 'quelque', 'chose']

Alignment:

[(0, 0), (1, 1), (2, 2)]

IBM Models

Statistical Machine Translation Models (IBM Models)

src	trg	prob
chien	dog	0.8
chien	cat	0.2
chat	cat	0.7

Scoring

BLEU: Gold standard

NSIT: Variation on BLEU

GLEU: Better for single sentences

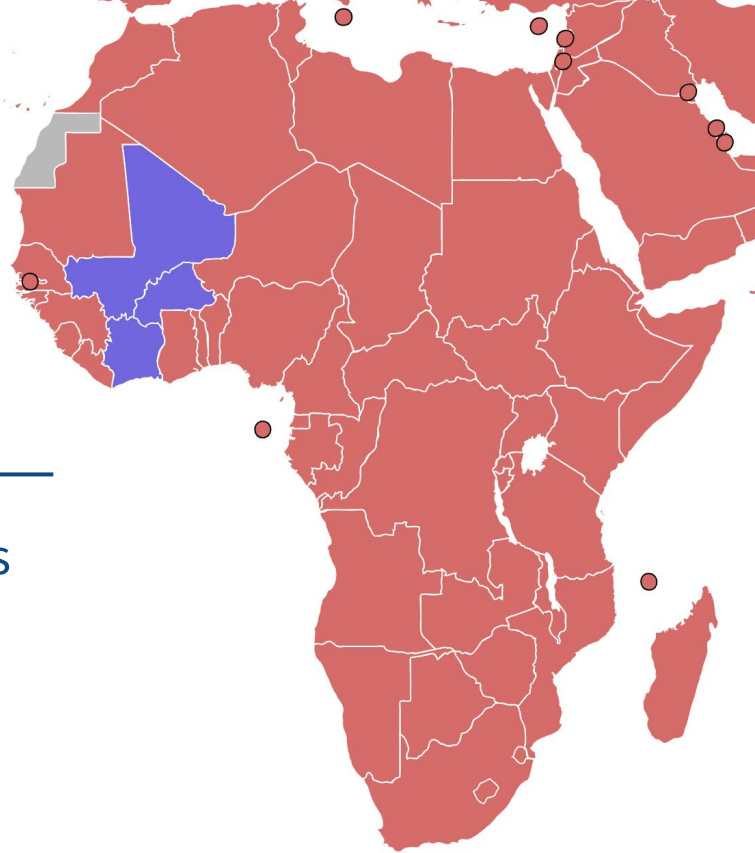
CHRF: Character Level

Case Study:
Dyula to French Translation

Objective:	Build a machine translation model from Dyula to French. This model will be used on platforms like discord.
Challenges:	Limited dataset, no GPU usage (CPU-only inference), and strict constraints on latency and memory.
Limited Dataset:	8000 sentence pairs
Inference Latency:	Optimize for speed by
CPU Usage and Memory Impact:	Need to use light weight models

Duyla

*Language in **West Africa** in **Ivory Coast**, **Burkina Faso** and **Mali***



(1) Dyula is closely related to Bambara and is part of the Mande language family
(2) It is mainly spoken and its written form isn't as standardized or widespread
(3) Lack of large parallel corpora

Example: *o bɛ bi bɔra fo gubeta*
Translation: *ah it's him he's ringing*

Case Study: Translation using NLTK

Steps	What was done	NLTK Tools
Preprocessing	- **Punctuation** + **lowercase**	-
Tokenization	- **Word** tokenizer	nltk.tokenize.word_tokenize
Alignment	- Dyula sentences were **reversed** (subject-verb-object -> subject-object-verb)	nltk.translate.AlignedSent
Model	**- IBM Model 2** **- Back Translation** was used to increase training dataset **- Alignment** was considered on **both directions** (source→target and target→source) and probabilities updated	nltk.translate.IBMModel2

Case Study: Translation using Transformers

Steps	What was done	Tools
Preprocessing	- **Punctuation** + **lowercase**	-
Tokenization	- Pretrained **NLLB** tokenizer	transfomer.NLLBTokenzier
Model	- **NLLB** pretrained model - **Pruned** model - **Quantized** model - Trained **Student Model** from Teacher	transformers.optimization.Adafactor transformers.AutoModelForSeq2Se1LM Ctranslate2 transformers.M2M100COnfig transformers.M2M100ForConditionalGeneration transformer.Seq2SeqTrainer transformer.Seq2SeqTrainingArgumenrs

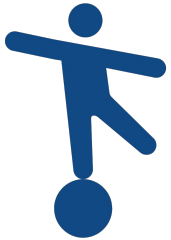
Case Study: Model Comparisons

	NLTK Model	Transformer Model
Results are rough benchmarks used for comparisons		

Case Study: Model Comparisons

	NLTK Model	Transformer Model
Training Time	**15 mins** on CPU (Apple M1, Total Number of Cores:8, Memory: 16 GB)	**8 hours 42 mins** on GPU (NVIDIA Tesla K80 with 12GB of VRAM)
Sentence GLEU (mean)	**5.3%**	**6.8%**
Memory Utilization	**55 % (2 GB)**	70% (2 GB)
Image Size (docker)	**555.44 MB**	2.45 **GB**
Prediction time (70 examples)	**CPU times: user 1.13 ms, sys: 1.72 ms, total: 2.85 ms** **Wall time: 1.88 ms**	CPU times: user 5.7 s, sys: 139 ms, total: 5.84 s Wall time: **7.85 s**

Final Thoughts



Translation is balance

Thomas Sowell ~ *“There are no solutions, only trade-offs.”*

Small Models, Big Impact

You can build meaningful models using simpler methods

The Power of NLTK

NLTK library, though simple, offers powerful tools

Always Evolving

While Transformer models dominate, there is still room for traditional approaches

Thank you

praelexis