TextAsCorpusRep

Multilingual Text As Corpus Repository for Machine Translation of Low-Resource Languages

Christian Schuler, Tramy Thi Tran, Deepesha Saurty, Anran Wang, Raman Ahmad, Seid Muhie Yimam

DDLitLab Student's Project
University of Hamburg

2024-04-17

Project Team

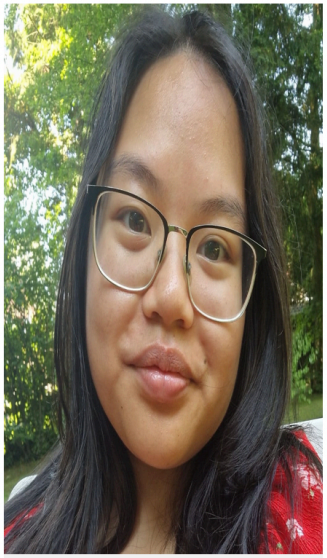
Christian Schuler

Master student in computer science at University of Hamburg.

With a passion for languages and colors, he aims to do a PhD. Additionally, he eats more pizza in one year than any of you in your lifetime!

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Tramy Thi Tran



Bachelor student in something like computer science at University of Hamburg.

She is not simply the head of our HR-department, she IS the entire department! XD #bestProjectCoordinatorInTown

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Deepesha Saurty

Bachelor student in computer science at University of Hamburg.

By now she went down the rabbit hole of science so much, that people actually heard her say "Algorithms are fun!"

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Anran Wang

Master student in computer science at Technical University Munich.

She is so awesome, that she did not start looking for a PhD position like most mortal beings do- the PhD position came falling out of the sky and had to apply to her first!

Raman Ahmad

Bachelor student in computer science at Hochschule für Angewandte Wissenschaften in Hamburg.

He is currently exploring various ways to advance the low-resource language Kurdish with a focus on the Kobani dialect.

HAW HAMBURG

Seid Muhie Yimam

Technical Lead at HCDS and Research Associate at Language Technology Group in Hamburg.

He researches low-resource languages, in particular related to the Amharic language, while he uses his spare time to mentor and supervise groups of students like noone else could! :D

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Our Idea

Translation Systems (Here Google Translate) Failing Languages

German	Ich mag die Farbe Grün, esse gerne Pizza, und meine Schwester wohnt in Österreich.
English	I like the color green, I like eating pizza, and my sister lives in Austria. ☺
Chinese	我喜欢绿色，我喜欢吃披萨，我姐姐住在奥地利
Chinese (Pinyin)	Wǒ xǐhuān lǜsè, wǒ xǐhuān chī pīsà, wǒ jiějiě zhù zài àodìlì. ☺
Vietnamese	Tôi thích màu xanh lá cây, tôi thích ăn pizza và chị gái tôi sống ở Áo.
	Tôi thích màu xanh, tôi thích ăn pizza và chị gái của tôi sống nước Áo. 😐
Kurdish	Ez ji rengê kesk hez dikim, ez ji xwarina pizza hez dikim, û xwişka min li Avusturya dijî.
Sorani / Kurmanji ?	Ez ji rengê kesk hez dikim, ez ji xwarina pîzayê hez dikim, û xwîşka min li Awistiryayê dijî 😐
Morisien	... ☹
Kobani	... ☹
~7000 more	... ☹

Lack of Standardization

Dialect Group	#Var	#ISO Codes	#Wiki (Articles)
Central Kurdish	13	2	53,856
Northern Kurdish	28	1	75,358
Southern Kurdish	13	1	0
Zazaki	10	3	41,811
Gorani	13	4	0

Number of language varieties (#Var) and corresponding ISO-Codes (#ISO) according to Glottolog [1], number of wikipedia articles (#Wiki) according to Wikimedia [2].

[1] http://glottolog.org
[2] https://meta.wikimedia.org/wiki/List_of_Wikipedias_by_language_group

Vietnamese

Name:

- Vietnamese (Tiếng Việt)

Vietnam: Country in Southeast Asia

- Prior to the alphabet reform, Vietnam used Chinese characters (Chữ Nôm)
- When Vietnam was conquered by the French, the Latin alphabet (Chữ Quốc Ngữ) systematically replaced the Chinese characters as the written language for everybody -> greatly increased literacy in Vietnam

Vietnamese: Three main dialect regions

- More than 80 million native speakers
- Various dialects, generally mutually intelligible
- Vocabulary influenced by Chinese and French

Kurdish Kobani

Name:

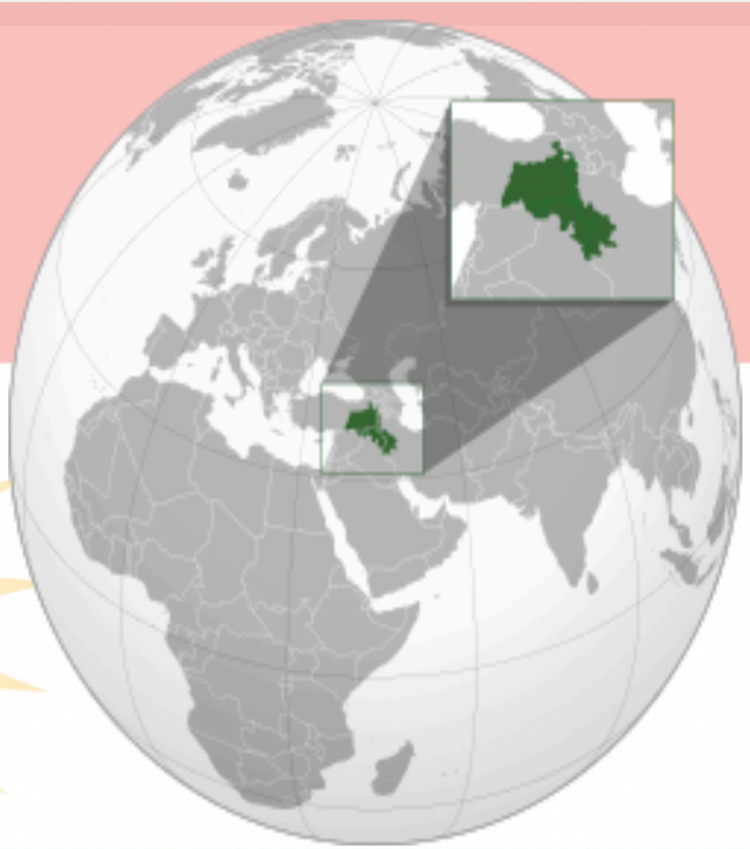
- Kurdish (Kurdî)
- Kurdish -> Kurmanji -> Southwestern -> Kobani

Kurdistan: Not being an officially recognized nation

- Scene of armed struggles since First World War
- No government that organizes or funds language research
- Population separated over nations and language spheres

Kurdish: Made up of a vast dialect continuum

- Approximately 25 million speakers
- Sorani (Central Kurdish) ~7 million Kurds
- Kurmanji (Northern Kurdish) ~15-20 million Kurds
- For a long time illegal to read & write



Mauritian Creole

Name:

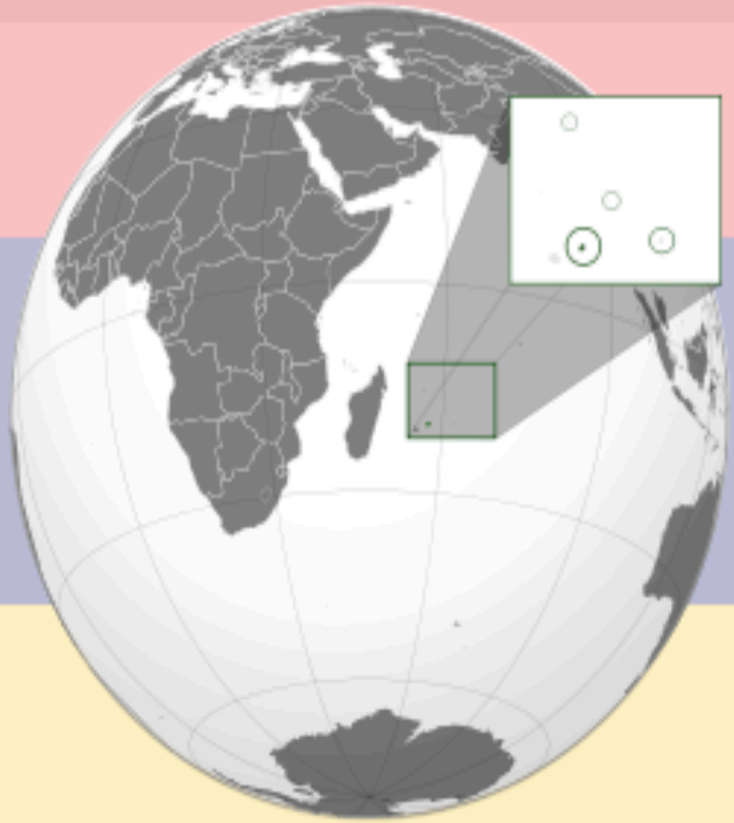
- Mauritian Creole or Morisien (formerly Morisyen)

Mauritius: An island nation and popular tourist spot

- Morisien is a young language influenced by many others
- Only recently formalized grammar and spelling

Morisien: A relatively small language community

- Approximately 1.3 million speakers
- French-based, with some English loan-words
- Different spellings and grammar rules



Chinese

Name:

- Chinese (simplified) (汉语)

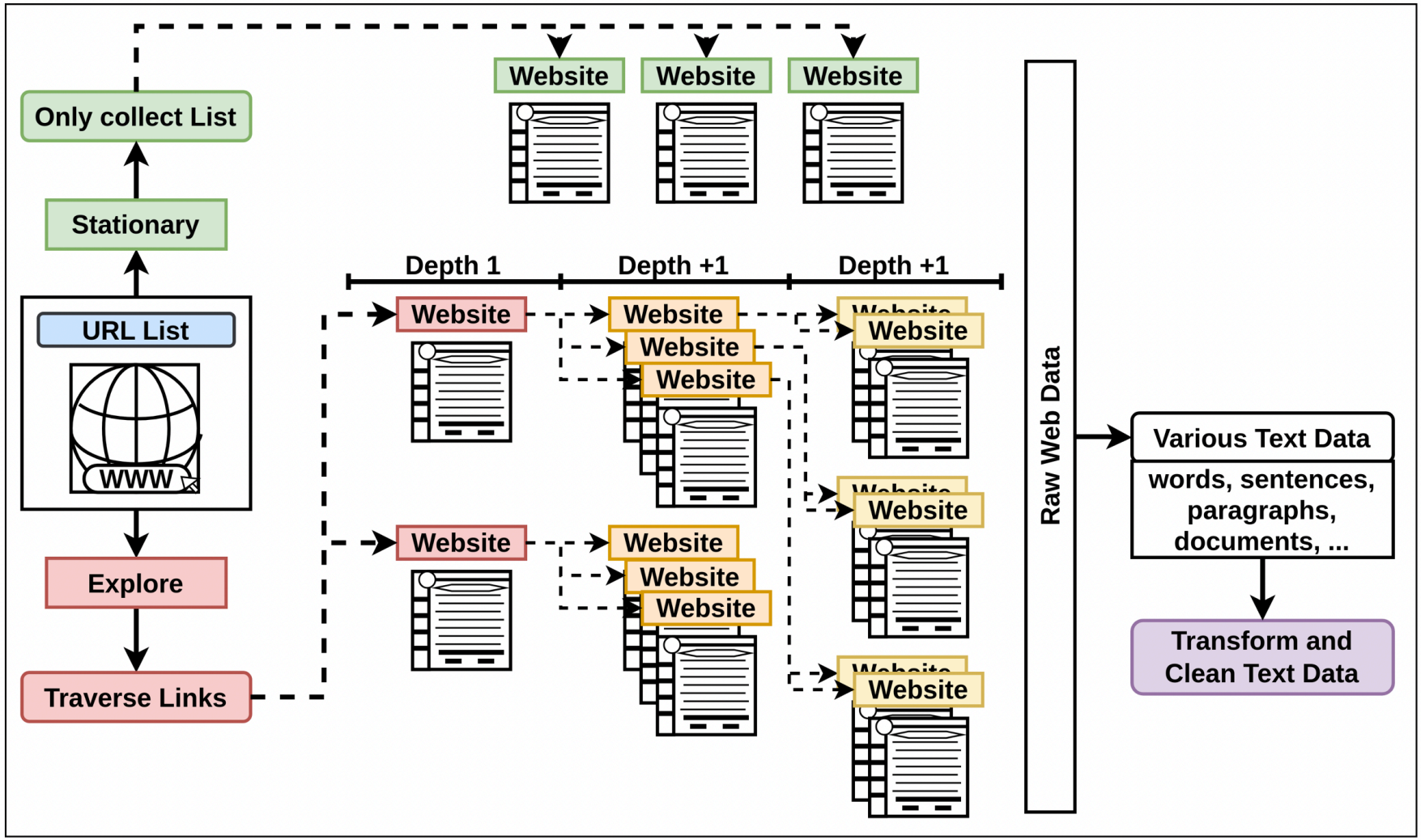
China: Country in East Asia exceeding 1.4 billion inhabitants

- Chinese is classified into more than 10 dialect groups
- Many of which are mutually unintelligible

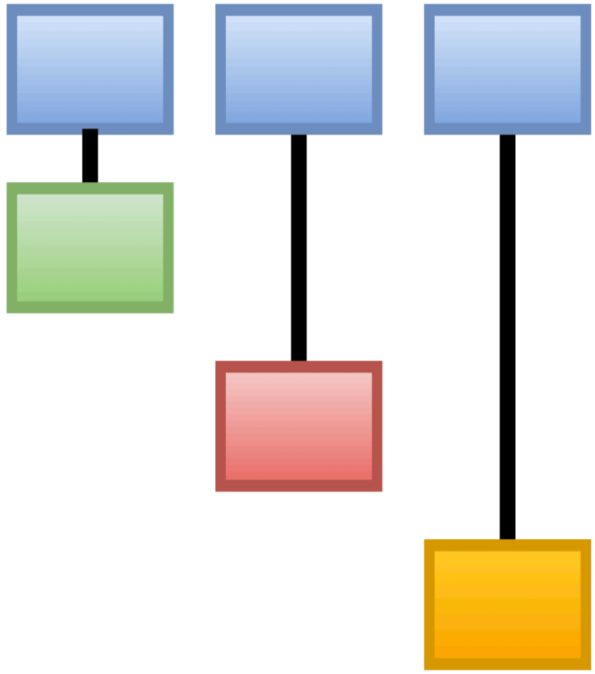
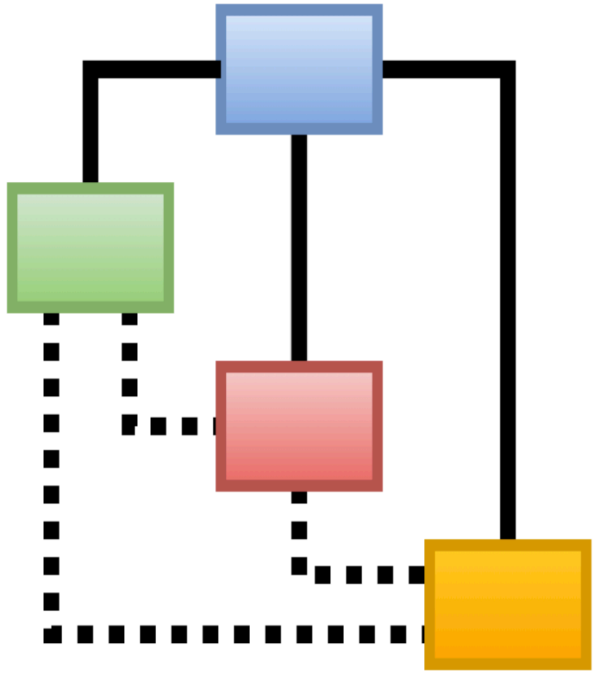
Chinese: A standardized writing system used for all variants

- Many low-resource variants, not even properly classified
- Standard Chinese though, provides lots of resources & tools
- Proposed as alternative to English-centric NLP research

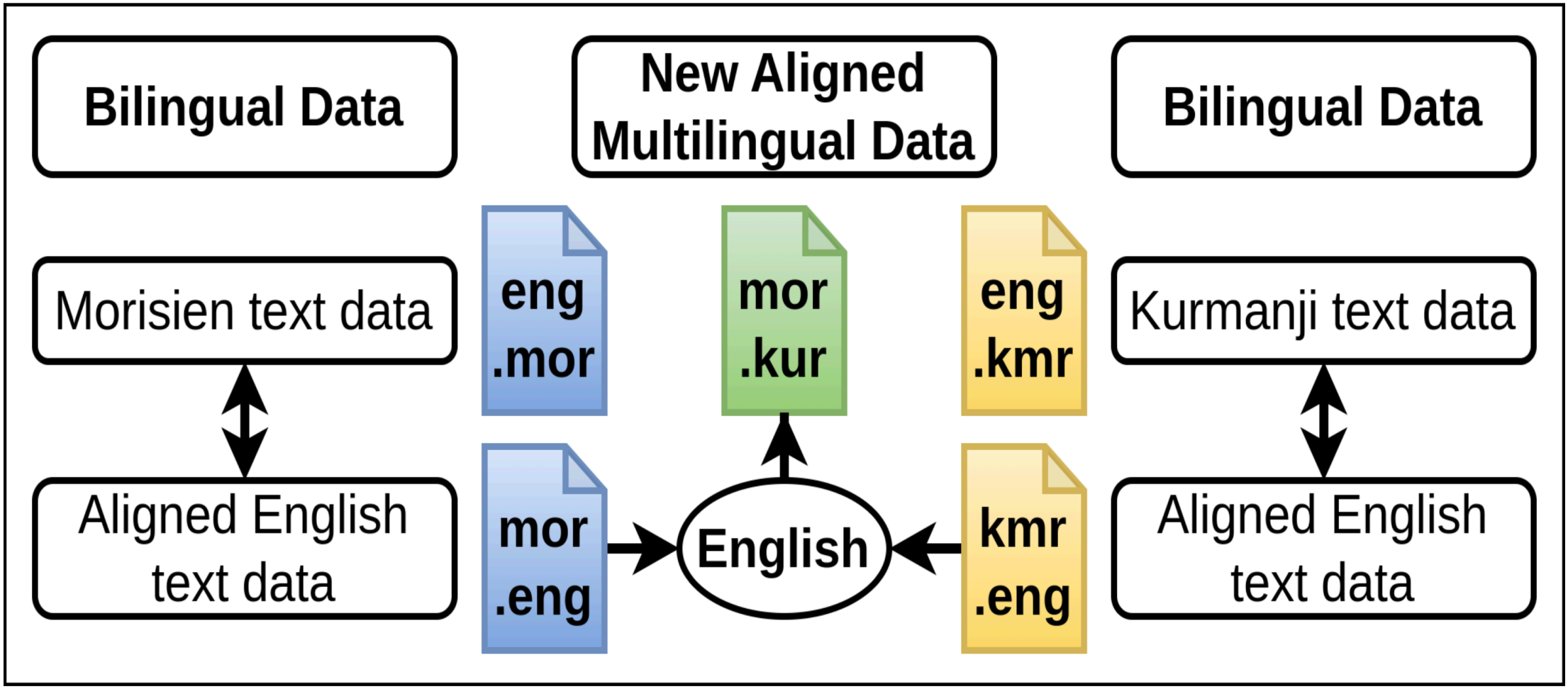
Collecting Text Data



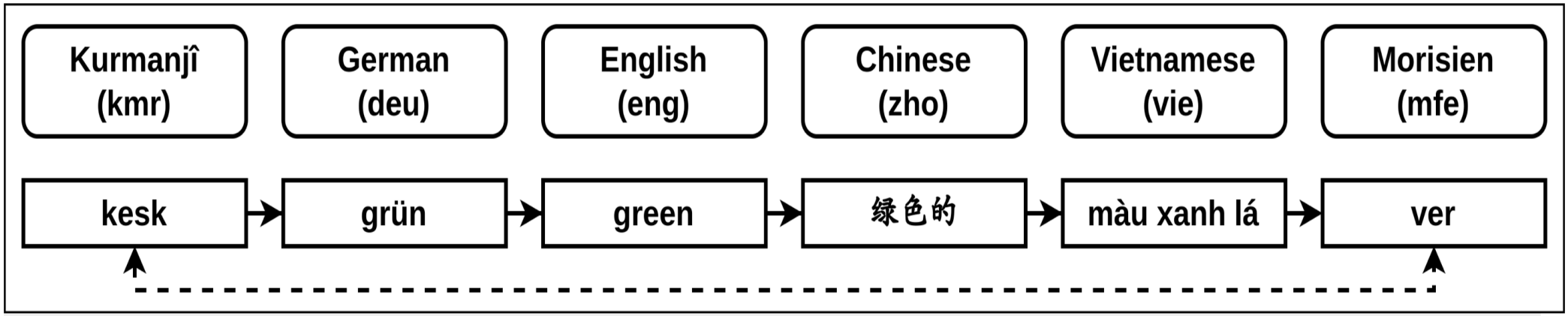
Bridging the Language Gap

Cross-Lingual Aligments	Directly Aligned	Indirectly Aligned
Language A		
Language B		
Language C		
Language D		

Use of Pivot Languages

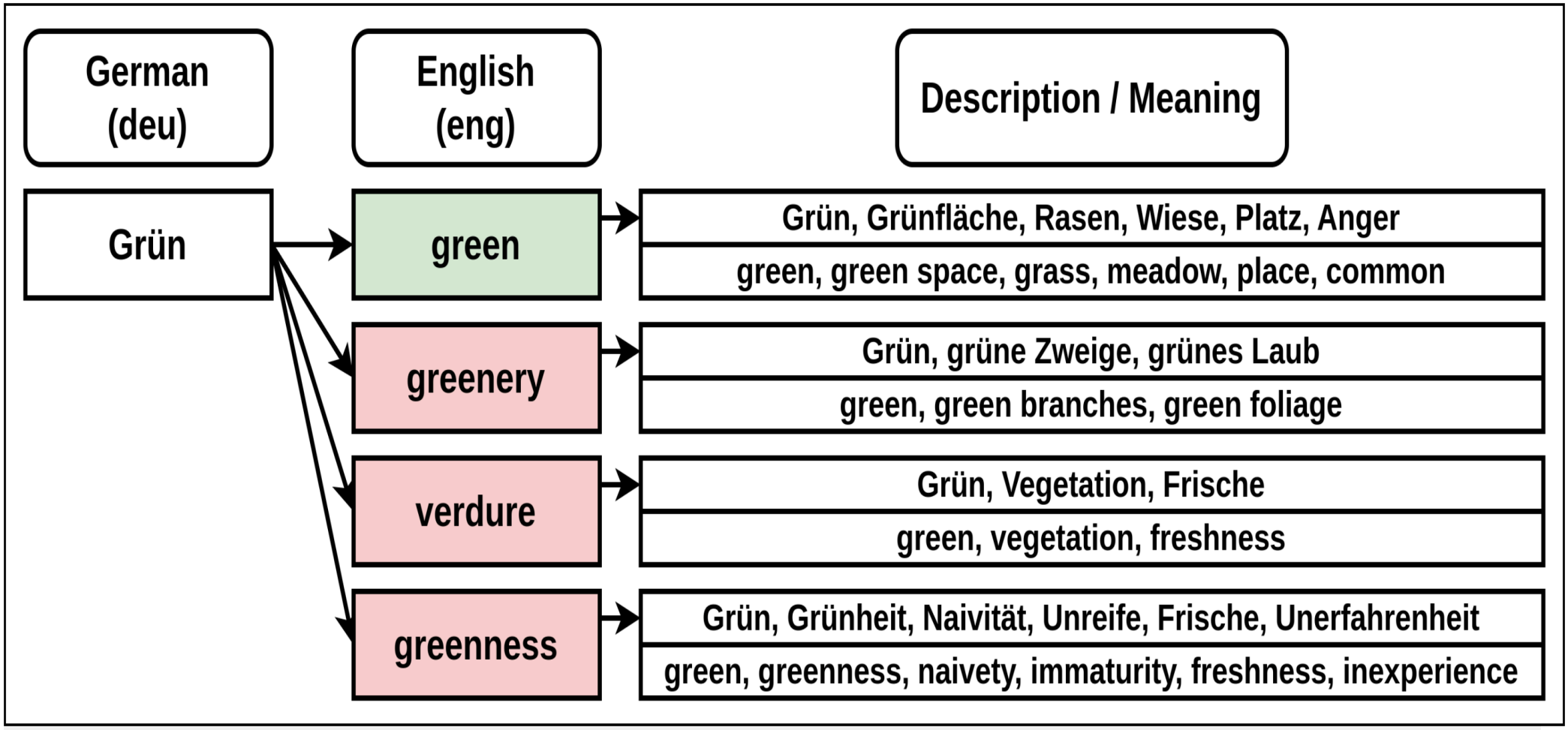


Language Ambiguities I



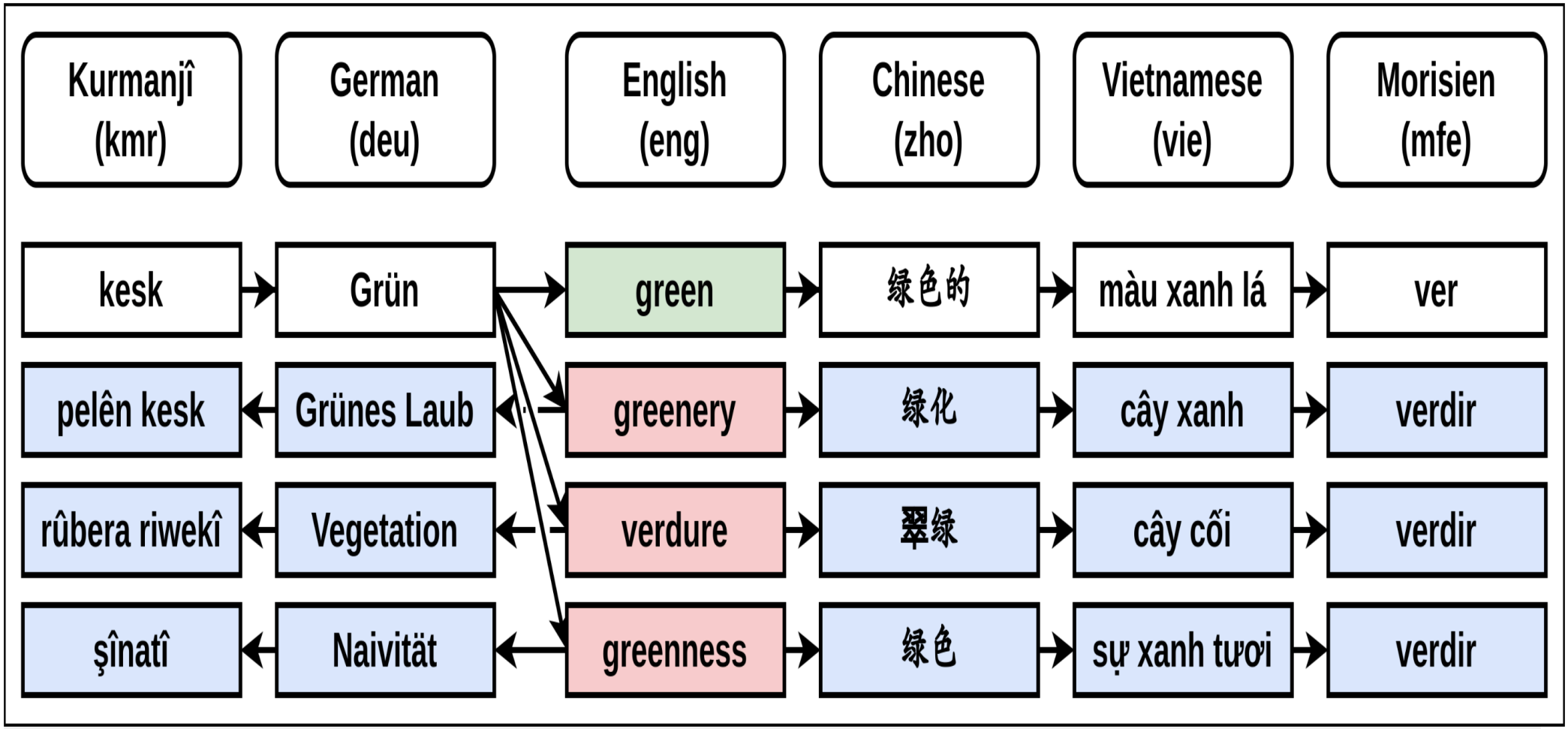
Finding new word alignments by traversing multiple cross-lingual alignments and assuming these relations to be transitive.

Language Ambiguities II



The word ambiguity during translation of the German Grün into the English green resulting in additional translation candidates.

Language Ambiguities III



Continuing on ambiguous translations (red) instead of aggregating to the most likely candidate (green) can enrich a dataset with additional labels across languages (blue).

Collecting Multilingual Text Data

Translating from English

No Language Left Behind (NLLB) is a first-of-its-kind, AI breakthrough project that open-sources models capable of delivering evaluated, high-quality translations directly between 200 languages— including low-resource languages like Asturian, Luganda, Urdu and more.

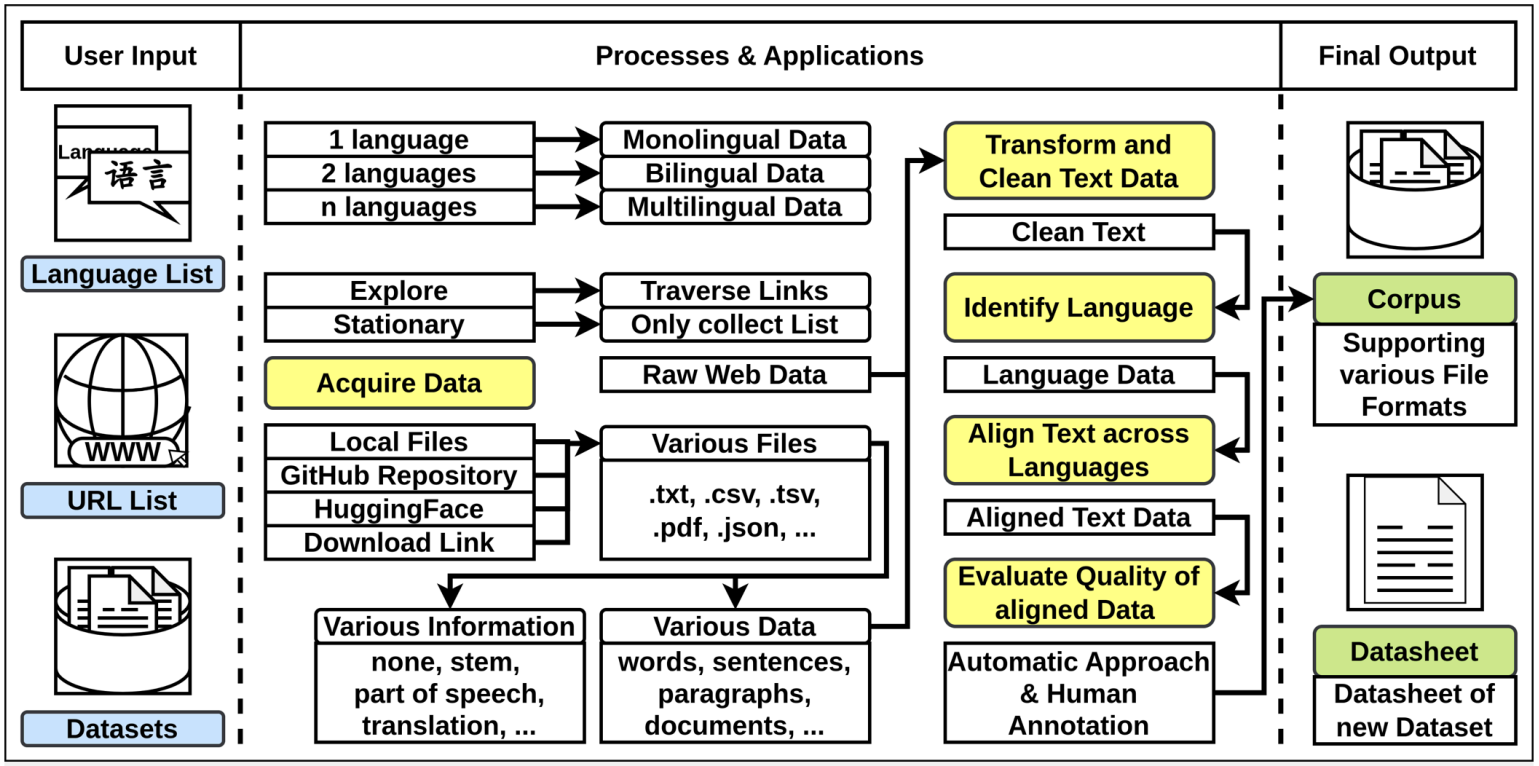
It aims to give people the opportunity to access and share web content in their native language, and communicate with anyone, anywhere, regardless of their language preferences.

Monolingual Descriptions

The Flickr30k dataset contains 31,000 images collected from Flickr, together with 5 reference sentences provided by human annotators.

Has already been extended to German, French, Czech, Chinese, and Ukrainian

Overview of Our Workflow



Our inputs in blue, main processing steps in yellow and resulting artifacts in green.

GlotLID

- Open source language identification tool
- Recognizes **more than 1600 languages**
  - Google Translate can recognize about 130 languages

- **Input:**

```
p1 = model.predict("Zordi mo ti al aste bann zafer dan soupermarse")
p2 = model.predict("Today, I went grocery shopping", 2)

print("\n")

print("Language: ", p1[0], " | Accuracy: ", p1[1])
print("Language: ", p2[0], " | Accuracy: ", p2[1])
```

- **Output:**

```
Language:  ('__label__mfe_Latn',)  | Accuracy:  [0.99980718]


Language:  ('__label__eng_Latn', '__label__pol_Latn')  | Accuracy:  [9.99628425e-01 1.76668356e-04]
```

Potato Annotation Tasks

Mô tả hình ảnh bằng tiếng Việt

test@test

••••••••••

LOGIN

Forgot Username / Password?

Login → Create your Account →

A. Projections are a subject of several pure mathematical fields, including differential geometry, projective geometry, and manifolds.
B. Rather, any mathematical function that transforms coordinates from the curved surface distinctly and smoothly to the plane is a projection.
C. The Earth and other large celestial bodies are generally better modeled as oblate spheroids, whereas small objects such as asteroids often have irregular shapes.

Please translate the 3 sentences from the above text one by one.

For A:

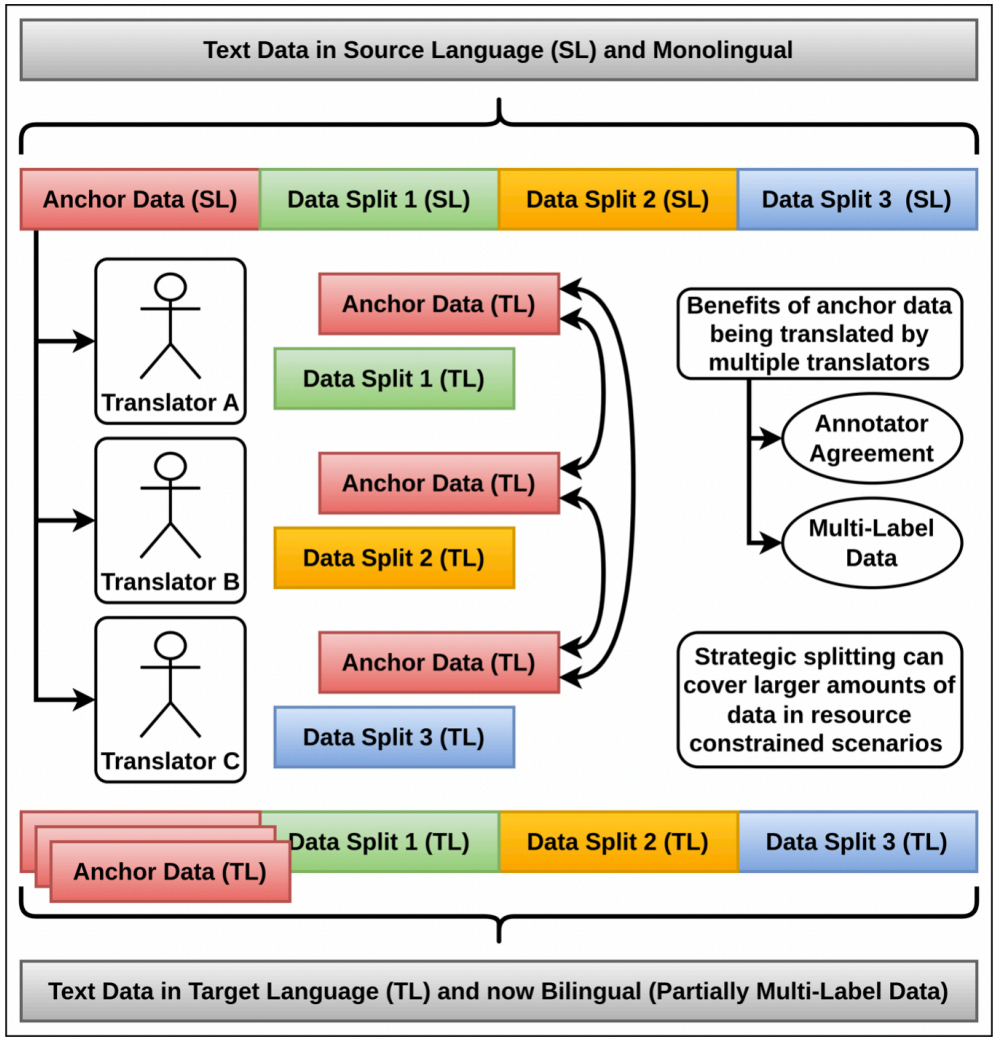
For B:

For C:

Hãy mô tả nội dung hình ảnh trong một câu và sử dụng tiếng Việt.

Sự miêu tả: Một người đàn ông mặc quần jean, tóc nâu và đeo kính chơi ghi-ta

Annotation via Anchor Items



Data Collected (NLLB-01)

- English (Original Text)
- "Lillian Diana Gish (October 14, 1893 – February 27, 1993) was an American actress, director and screenwriter."

- German (MT via GoogleTranslate and ChatGPT)
- "Lillian Diana Gish (14. Oktober 1893 – 27. Februar 1993) war eine US-amerikanische Schauspielerin, Regisseurin und Drehbuchautorin."
- "Lillian Diana Gish (14. Oktober 1893 – 27. Februar 1993) war eine amerikanische Schauspielerin, Regisseurin und Drehbuchautorin."

- Morisien
- "Lillian Diana Gish (inn ne 14 Oktob 1893 – inn mor 27 Fevrie 1993) li ti enn akter, enn realizater ek enn senaris Amerikenn."
- "Lillian Diana Gish (14 Oktob 1893 - 27 Fevriye 1993) ti enn aktris, realizatris ek senarist amerikenn."

- "Vietnamese"
- "Lillian Diana Gish, sinh ngày 14 tháng 10 năm 1893 và mất ngày 27 tháng 2 năm 1993, là một nữ diễn viên, đạo diễn và biên kịch người Mỹ."
- "Lillian Diana Gish (14/10/1893 – 27/2/1993) là một nữ diễn viên, đạo diễn, biên kịch người Mỹ."
- "Lillian Diana Gish (14/10/1893 – 27/02/1993) là một nữ diễn viên, đạo diễn và nhà biên kịch người Mỹ."

- "Kobani"
- "Lillian Diana Gish (14'î mehê 10'an 1893 - 27'î mehê didîyan 1993) lîstikvan, derhêner û sînaronivîseke Emrîkî bû."
- "Lilian Diana Gish (14î meha dehan, 1893 - 27î meha didiyan , 1993) lîstikvan, derhêner, û senarîsteke Emrîkî bû."

- Chinese
- "莉莲·戴安娜·吉什（1893年10月14日-1993年2月27日），美国女演员、导演和编剧。"
- "莉莲·戴安娜·吉什（1893年10月14日－1993年2月27日）是一位美国女演员、导演和编剧。"
- "莉莉安·黛安娜·吉许（1893年十月十四日-1993年二月27日）是一个美国演员，导演和编剧。"

Data Collected (NLLB-02)

- English (Original Text)
- "The seventeen-year-old Lillian traveled to Shawnee, Oklahoma, where James's brother Alfred Grant Gish and his wife, Maude, lived."

- German (MT via GoogleTranslate and ChatGPT)
- "Die siebzehnjährige Lillian reiste nach Shawnee, Oklahoma, wo James‘ Bruder Alfred Grant Gish und seine Frau Maude lebten."
- "Die siebzehnjährige Lillian reiste nach Shawnee, Oklahoma, wo James' Bruder Alfred Grant Gish und seine Frau, Maude, lebten."

- Morisien
- "Laz diset-an, Lillian ti al rest dan lavil Shawnee an Oklahoma (Etazini), kot frer so papa James, ki apele Alfred Grant Gish ek so madam, Maude."
- "Lillian, ki ena diset an, ti al Shawnee, dan Oklahoma, kot frer James, Alfred Grant Gish ek so madam, Maude ti viv."

- "Vietnamese"
- "Ở tuổi mười bảy, Lillian đến Shawnee, Oklahoma, nơi anh trai của James, Alfred Grant Gish và vợ Maude sinh sống."
- "Lillian, khi đó 17 tuổi, đã đến thành phố Shawnee thuộc tiểu bang Oklahoma, nơi anh trai của James là Alfred Grant Gish và vợ của ông ấy sinh sống."
- "Cô nàng Lillian mười bảy tuổi đã đến Shawnee, Oklahoma, tại đây có người anh trai của James là Alfred Grant Gish và vợ Maude sinh sống."

- "Kobani"
- "Lillian 17-salî çû Shawnee, Oklahoma, cihê ku birayê James Alfred Grant Gish û jina wî, Maude, lê dijîyan."
- "Lilliana hevde salîn çû Shawnee, Oklahoma, cê ku birê James Alfred Grant Gish û jina xwe lê dijiyan."

- Chinese
- "17岁的莉莲前往俄克拉何马州的肖尼市，詹姆斯的哥哥阿尔弗雷德·格兰特·吉什和他的妻子莫德就住在那里。"
- "17岁的莉莲前往俄克拉荷马州的肖尼市，詹姆斯的哥哥Alfred Grant Gish和他的妻子Maude住在那里。"
- "17岁的莉莉安去了奥克拉荷马州的肖尼，在那里住过詹姆斯的兄弟阿尔弗雷德·格兰特·吉许和他的妻子莫德。"

Data Collected (Flikr30k-01)

- Kobani
- "mêrkek bi kombrêsê erdê xerab dikê"

- Morisien
- "enn missier p craze simin avec so machine craz roche"

- Vietnamese
- "Một công nhân xây dựng đang làm việc với một búa khoan."
- "Anh ấy đang phá dỡ đường"

- Chinese
- "一名工人正在拆除房屋前面的水泥路面"

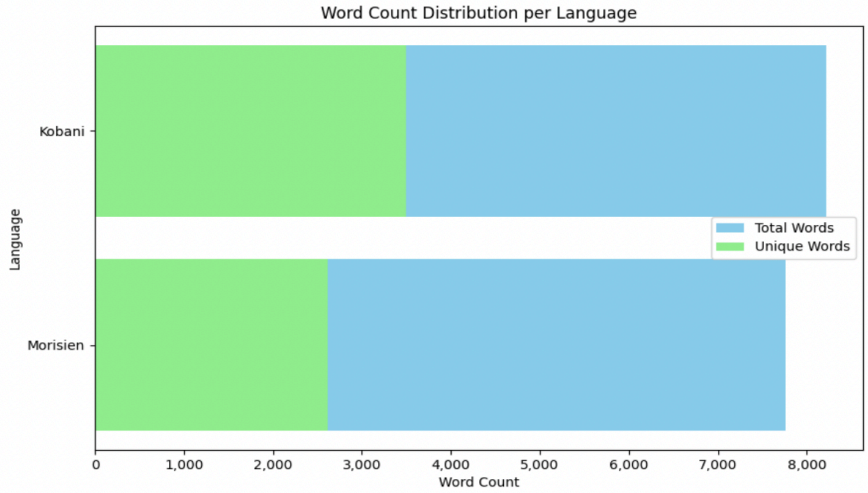
Data Collected (Flikr30k-02)

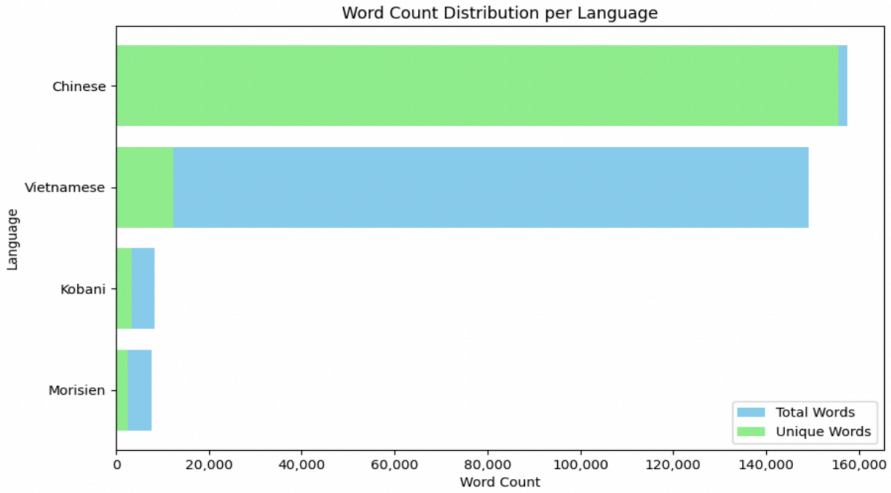
- Kobani
- "Qîzkek û mêrekî li baxçakî ne, li ber qîzikê hespek hê û li paş merik jî yek hê. Qîzik û hespî xwe li ber êr in."

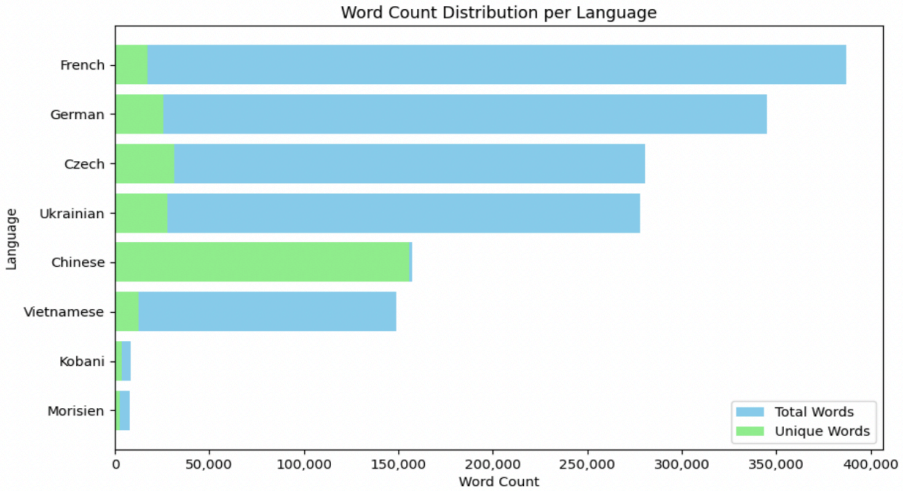
- Vietnamese
- "Cô gái đang chăn ngựa"
- "Trong trang là hình ảnh cô bé đang dắt 1 chú ngựa và phía đằng xa có lẻ là 1 huấn luyện viên của cô ấy thì phải."
- "Cô gái mặc áo đen đang đứng cầm dây cương ngựa, cả hai đứng bên cạnh một bếp lữa."

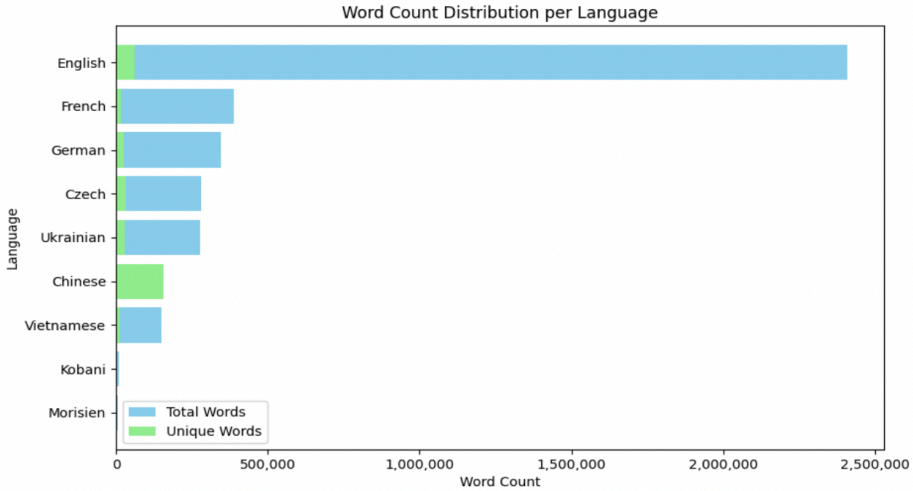
- Chinese
- "因为年轻女士牵着马站在火堆前"

Collected Data









Insights Gained

Doing this kind of work is [chết tiệt] hard!

Introducing CRAMT - A Participatory Research Approach

To enable others, even if

- less tech-savvy,
- new to NLP, or
- no idea where to start

CRAMT: Cross-Lingual Resource Aggregation of Low-Resource Machine Translation and Metadata

Christian Schuler[1], Tramy Thi Tran[1], Deepesha Saurty[1], Anran Wang[2], Raman Ahmad[3], Seid Muhie Yimam[1]

[1]Universität Hamburg [2]Technische Universität München [3]Hochschule für Angewandte Wissenschaften Hamburg

{christianschuler8989, raman.ahmad2022}@gmail.com, anran.wang@tum.de, {tramy.thi.tran, deepesha.saurty}@studium.uni-hamburg.de, seid.muhie.yimam@uni-hamburg.de

Introduction

This work addresses the issue of scant text data for **Machine Translation** of **low-resource languages** by introducing a **corpus creation tool**. This easy-to-use tool enables the creation of multilingual aligned text data for extremely low-resource languages. By including an annotation schema utilizing monolingual native speakers, even aggregated data of **zero-resourced languages** can be evaluated, resulting in higher quality datasets for these languages.

Motivation

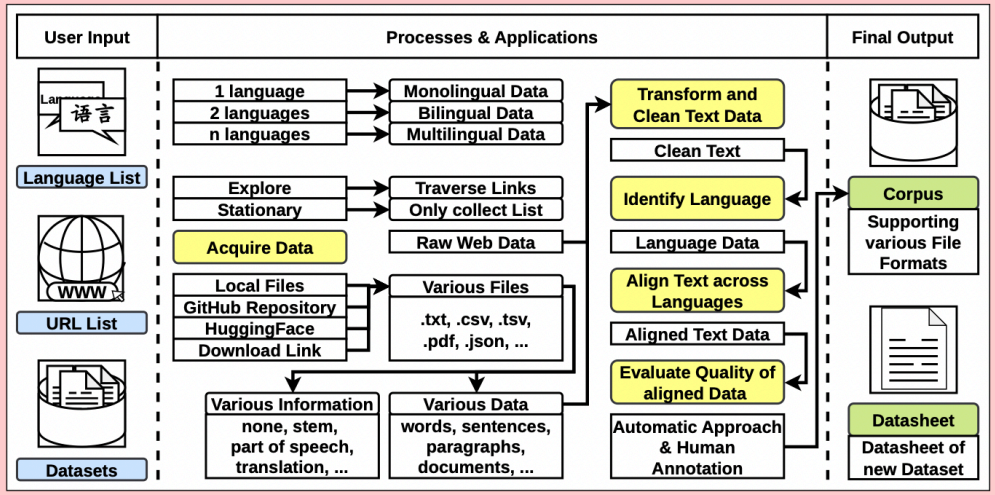
Google Translate: Translating German text

German	Ich mag die Farbe Grün, esse gerne Pizza, und meine Schwester wohnt in Österreich.
English	I like the color green, I like eating pizza, and my sister lives in Austria.
Chinese	我喜欢绿色，我喜欢吃披萨，我姐姐住在奥地利
Chinese (Pinyin)	Wǒ xǐhuān lǜsè, wǒ xǐhuān chī pīsà, wǒ jiějiě zhù zài àodìlì.
Vietnamese	Tôi thích màu xanh lá cây, tôi thích ăn pizza và chị gái tôi sống ở Áo.
	Tôi thích màu xanh, tôi thích ăn pizza và chị gái của tôi sống nước Áo.
Kurdish (Sorani / Kurmanji)	Ez ji rengê kesk hez dikim, ez ji xwarina pizzê hez dikim, û xwîska min li Avusturya dijî.
	Ez ji rengê kesk hez dikim, ez ji xwarina pîzayê hez dikim, û xwîska min li Awistiryayê dijî
Morisien	... Kobani ... ~7000 more ...

Lower performance of translation systems can often be linked to a severe lack of data for specific

Design & Development

The following graph shows how the toolkit is used: The user inputs their data in various forms, or provide URLs directing to the desired data, which then are preprocesed, cleaned, aligned, and evaluated, to finally produce a corpus supporting various formats, with a datasheet describing the new dataset.



CRAMT tool: User inputs in blue, main processing steps in yellow and resulting artifacts in green

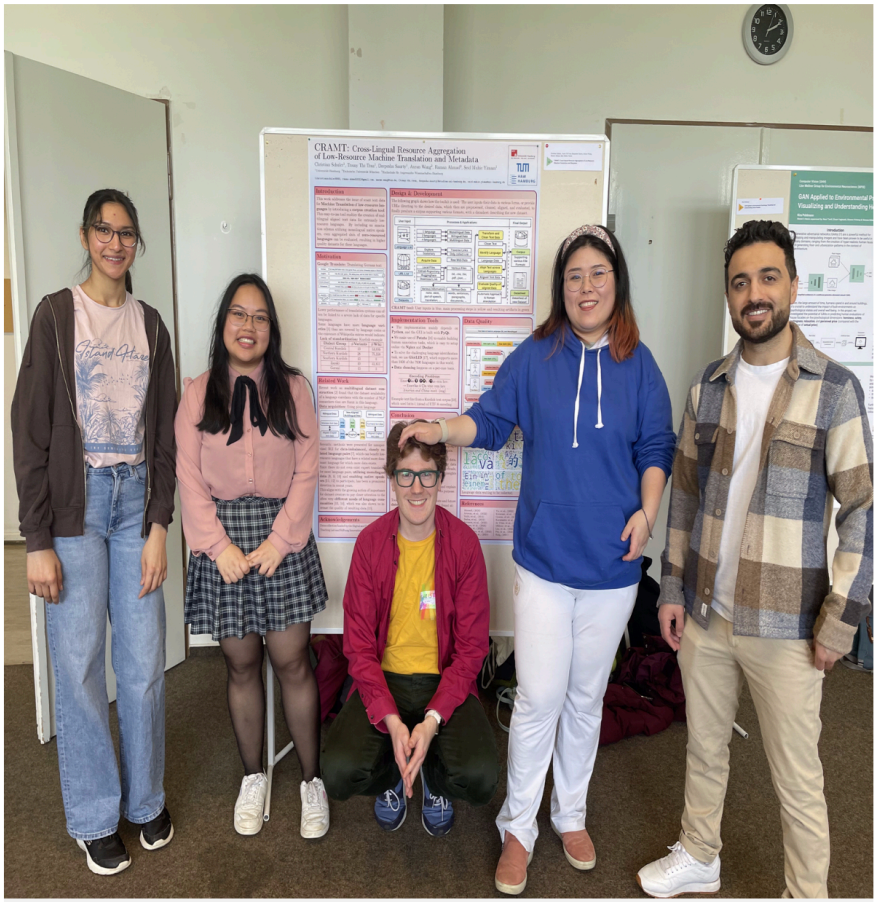
CRAMT at Workshop on Building and Using Comparable Corpora (LREC-COLING)

Close, but no cigar!
Additional to the reviews we received very encouraging words though:

The selection process was very competitive. Due to time and space limitations, we could only choose a limited number of the submitted papers to appear on the program.
Nonetheless, I still hope you can attend the conference.
Additionally, given the interest found in your project,
we encourage you to submit an updated version of your paper next year.

We are happy that you found the reviewers' comments useful.
As mentioned in the specific word added to the notification message,
we hope you can submit an updated paper next year!

CRAMT at EXPO-2024



Poster presentation to faculty, students, pupils.

CRAMT was awarded second place.

Language Data waiting to be collected

Kurmanji

Morisien

Vietnamese

Chinese

German

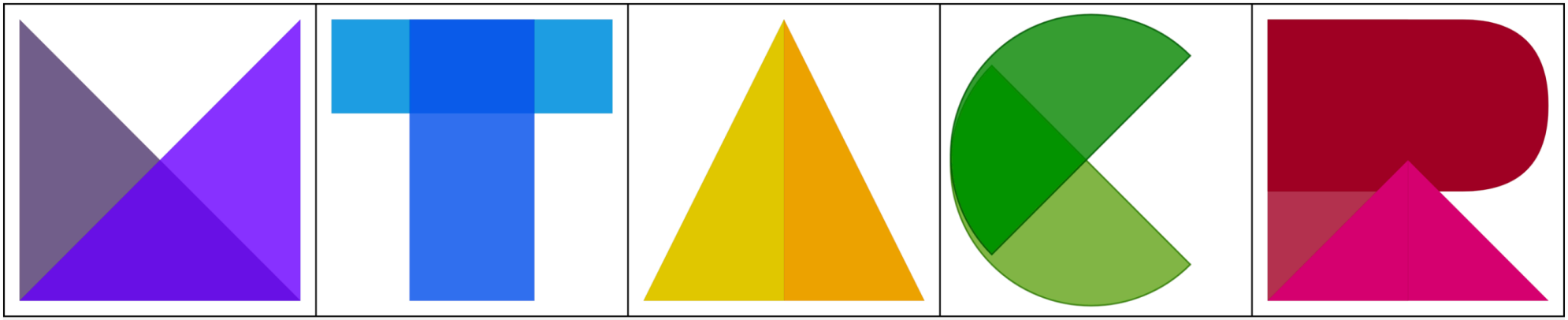
English

Let There Be Cake !

Conceptualization of future conference attendence.

- 1. CRAMT tool demo-paper at ACL or EMNLP(?)
- 2. CRAMT tool research-paper once we have
  - some more results and
  - the time to do more evaluations
- 3. MTACR corpus research-paper.

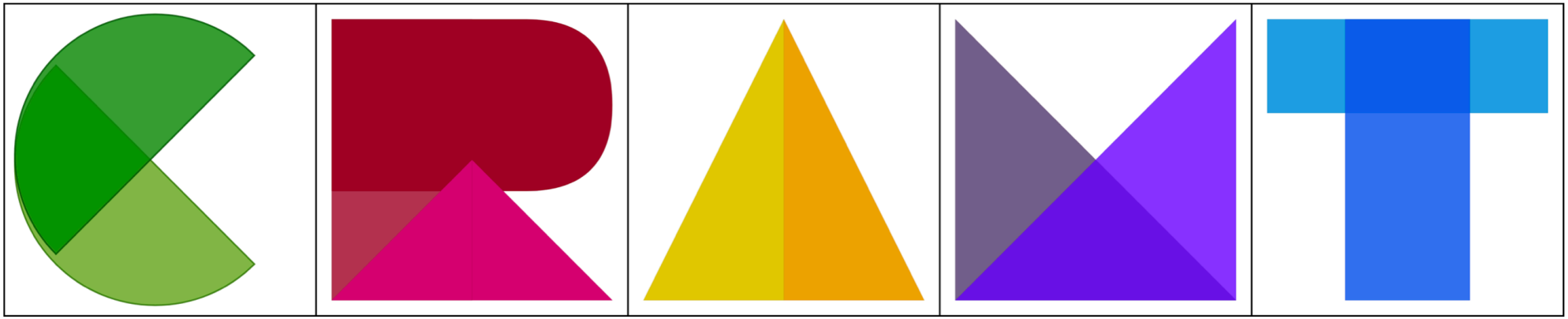
Main Project: TextAsCorpusRep



Multilingual Text As Corpus Repository for Machine Translation of Low-Resource Language

- Text Corpus for
  - Kurdish Kobani, Morisien Creole, Vietnamese, Chinese
- Novel Data
  - via scraping the internet + cleaning
  - via translations from English
  - via image descriptions by native speakers
- Previous Data
  - from available mono- and multi-lingual datasets

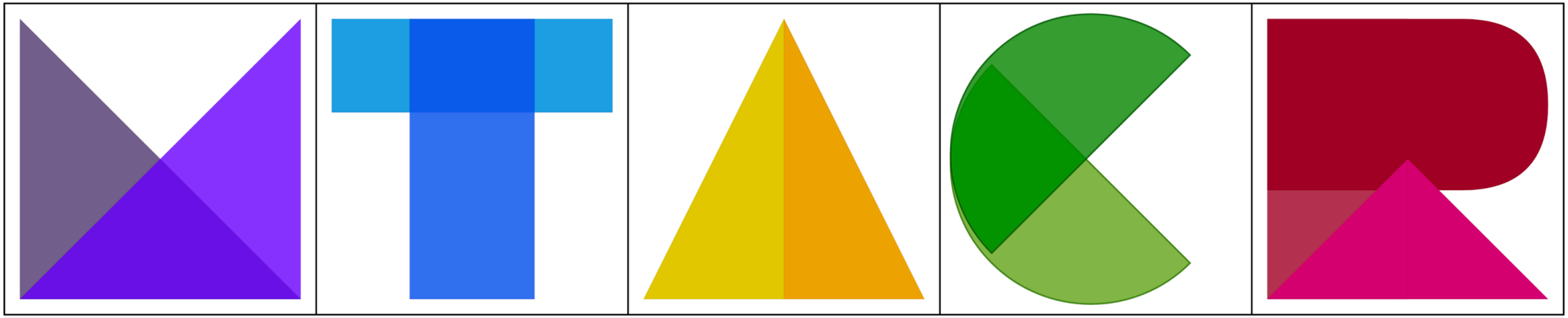
Auxiliary Project: CRAMT



Cross-Lingual Resource Aggregation of Low-Resource Machine Translation and Metadata

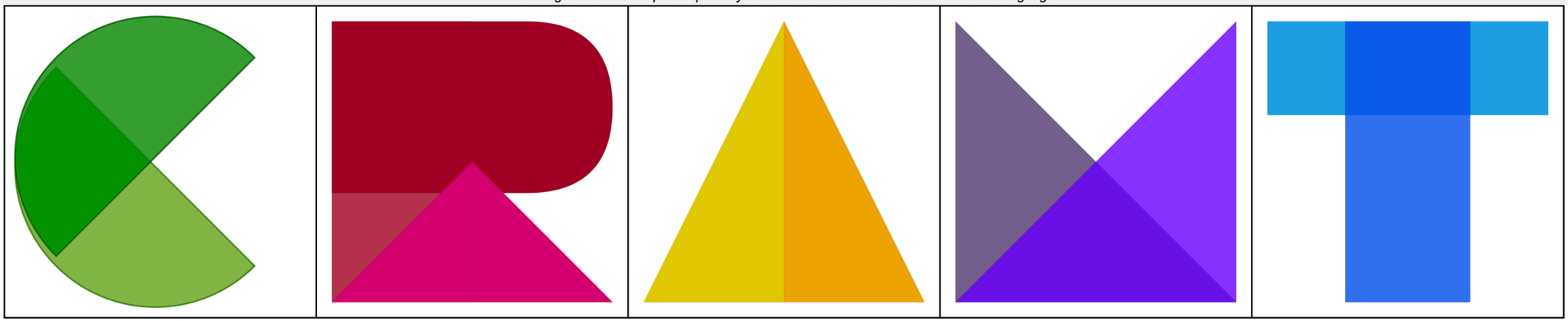
- 1. Text Corpus for specific target languages aiming to provide new aligned text data.
- 2. Analysis of the collected and aligned data.
  - Some help to get a quick idea of the data distribution such as generated word clouds,
  - but also reports to provide deeper insights about the data.
- 3. Datasheet that can represent and explain the newly created dataset and its purpose.

The End



Main Project: TextAsCorpusRep

Multilingual Text As Corpus Repository for Machine Translation of Low-Resource Language



Auxiliary Project: CRAMT

Cross-Lingual Resource Aggregation of Low-Resource Machine Translation and Metadata

References

MTACR-Related (TODO)

- Author, (Date)
- Author et al., (Date)

CRAMT-Poster-Related

- Ahmadi, (2020)
- Yu, et al., (2022)
- Artetxe, et al., (2022)
- Kreutzer, et al., (2022)
- Vulić, et al., (2013)
- Gouws, et al., (2016)
- Bafna, et al., (2023)
- Karakante, et al., (2018)
- Reimers, et al., (2020)
- de Vries, et al., (2021)
- Yimam, et al., (2020)
- Millour, et al., (2020)
- Lent, et al., (2022)
- Liu, et al., (2022)
- Cahyawijaya, et al., (2023)
- Pei, et al., (2022)
- Kargaran, et al., (2023)
- Haig, (2001)

CRAMT-Paper-Related (TODO)

- Author, (Date)
- Author et al., (Date)