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

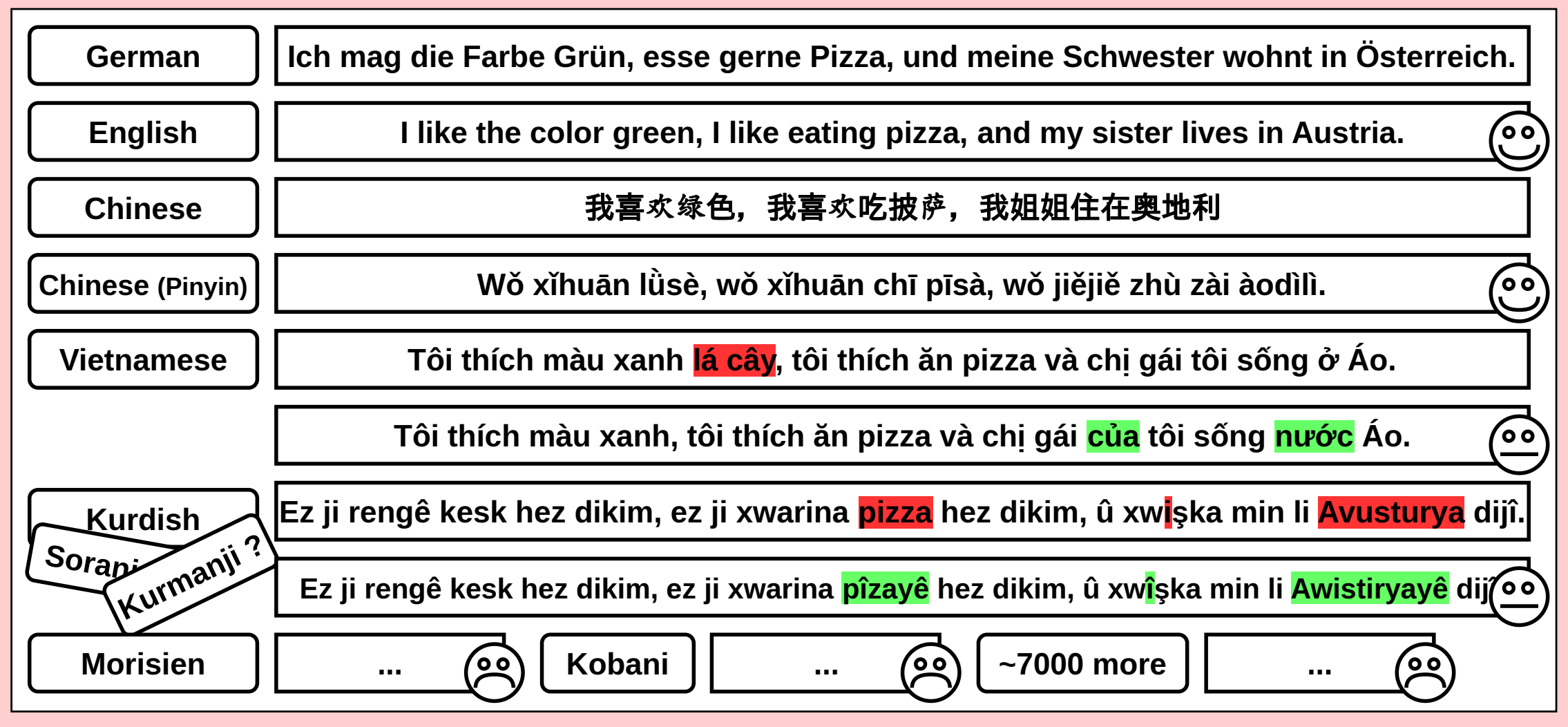
{christianschuler8989, raman.ahmad2022}@gmail.com, anran.wang@tum.de, {tramy.thi.tran, deepesha.saurty}@studium.uni-hamburg.de, seid.muhie.yimam@uni-hamburg.de

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 pizza hez dikim, û xwîşka min li Avusturya dijî.

Ez ji rengê kesk hez dikim, ez ji xwarina pîzayê hez dikim, û xwîşka min li Awistiryayê dijî.

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

Some languages have more **language varieties** [1] than are covered by language codes or the existence of Wikipedia entries would indicate.

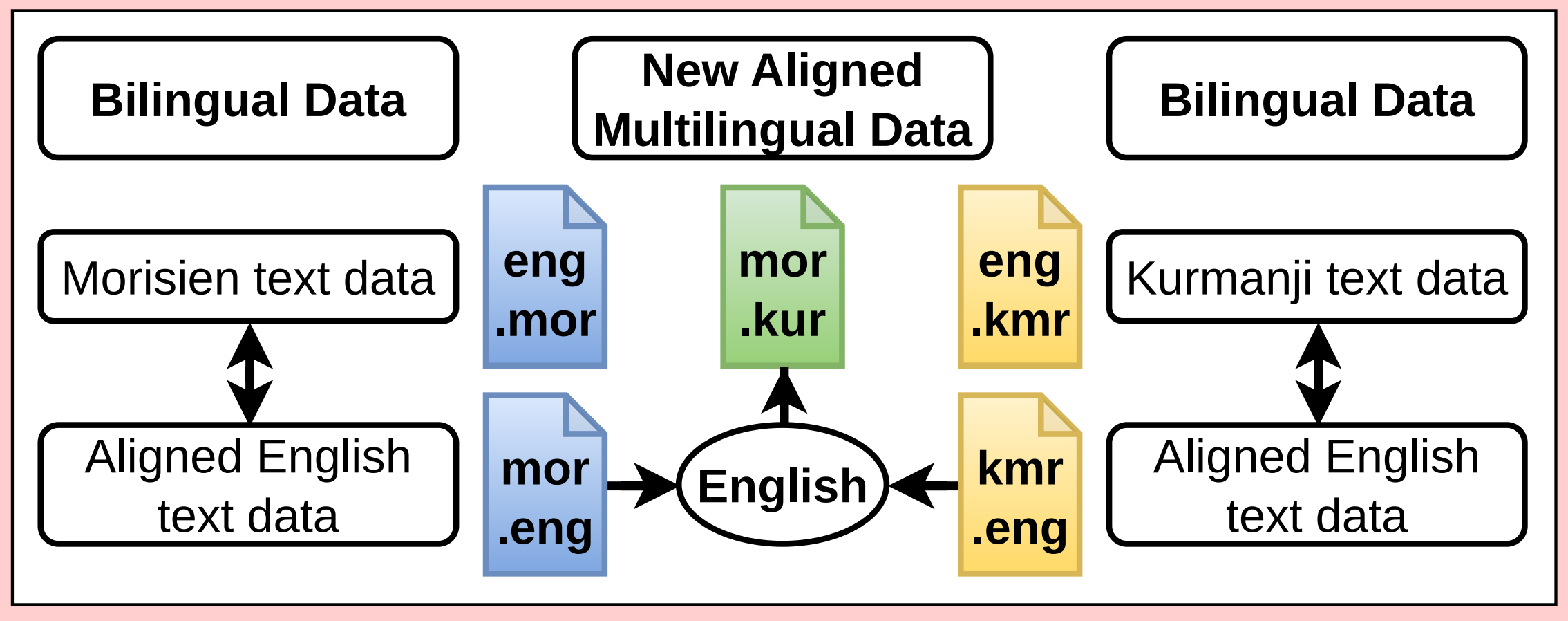
Lack of standardization: Kurdish example

Dialect Group	#Variants	#Wiki
Central Kurdish	13	53,856
Northern Kurdish	28	75,358
Southern Kurdish	13	0
Zazaki	10	41,811
Gorani	13	0

Related Work

Recent work on **multilingual dataset construction** [2] found that the dataset availability of a language correlates with the number of NLP researchers that are fluent in this language.

Data acquisition: Using pivot language



Recently, methods were presented for unsupervised BLI for **data-imbalanced, closely related language pairs** [7], which can benefit low-resource languages that have a related more dominant language for which more data exists.

Since there do not even exist expert translators for most language pairs, **utilizing monolingual data** [8, 9, 10] and **enabling native speakers** [11, 12] to participate, has been a prominent direction in recent years.

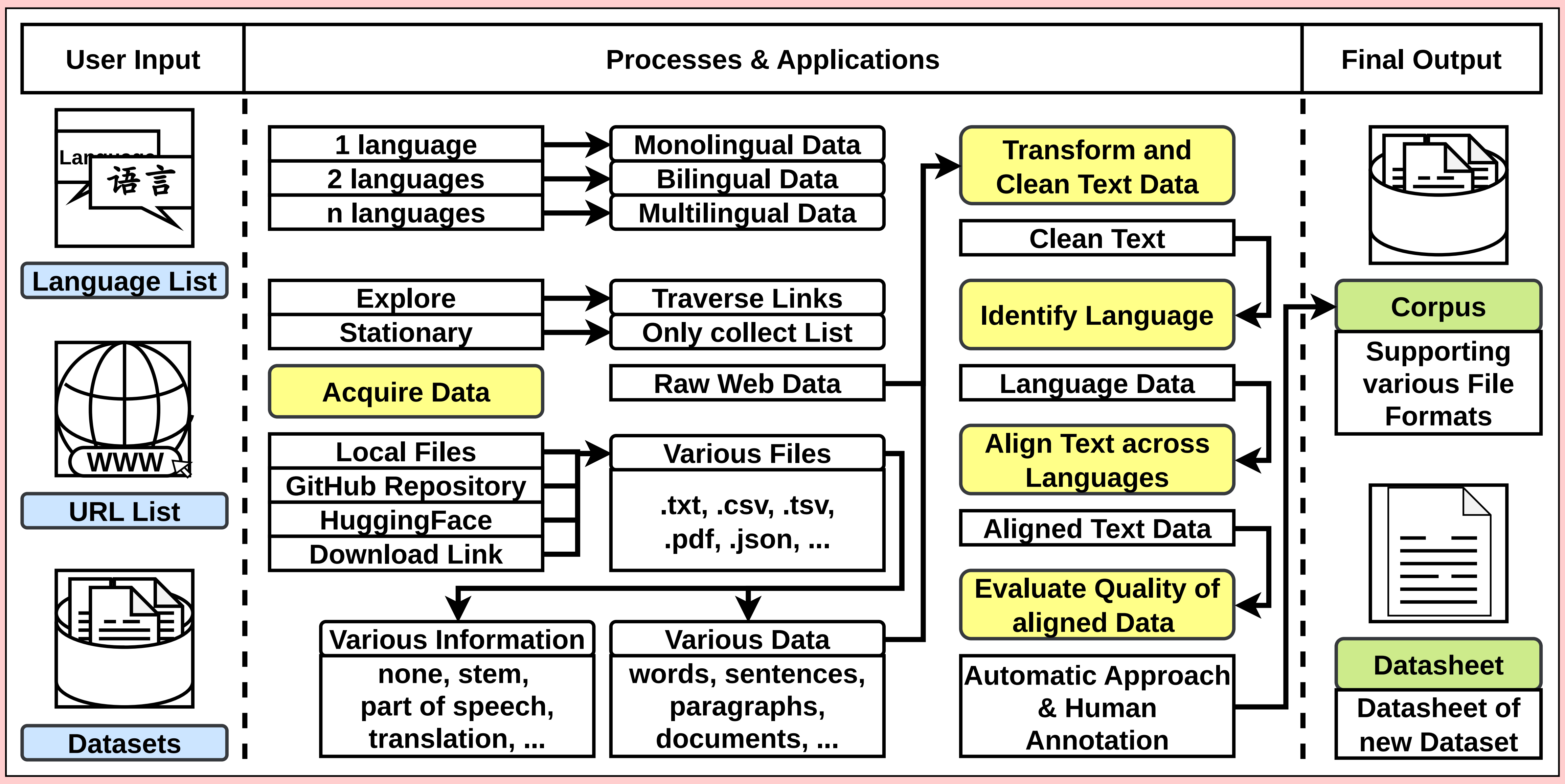
This aligns with the growing notion of importance for dataset creators to pay closer attention to the often very **different needs of language communities** [13, 14], which was also shown to increase the quality of resulting data [15].

Acknowledgements

Data collection funded by the Digital and Data Literacy in Teaching Lab and Stiftung Innovation in der Hochschullehre.

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

Implementation Tools

- The implementation mainly depends on **Python**, and the GUI is built with **PyQt**.
- We make use of **Potato** [16] to enable building human annotation tasks, which is easy to setup online via **Nginx** and **Docker**.
- To solve the challenging language identification task, we use **GlotLID** [17], which supports more than 1600 of the 7000 languages in this world.
- **Data cleaning** happens on a per-case basis.

Encoding Problems
Emer�ka � ��n t�ne cem hev.
→ Emerîka û Çîn têne cem hev.
America and China meet. (eng)

Example text line from a Kurdish text corpus [18], which used latin-1 intead of UTF-8 encoding.

Conclusion

This work resulted in **CRAMT**, a toolkit to collect data for low-resource languages, easily usable by MT research communities (experts and non-experts alike). The toolkit's results are three-fold:

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.

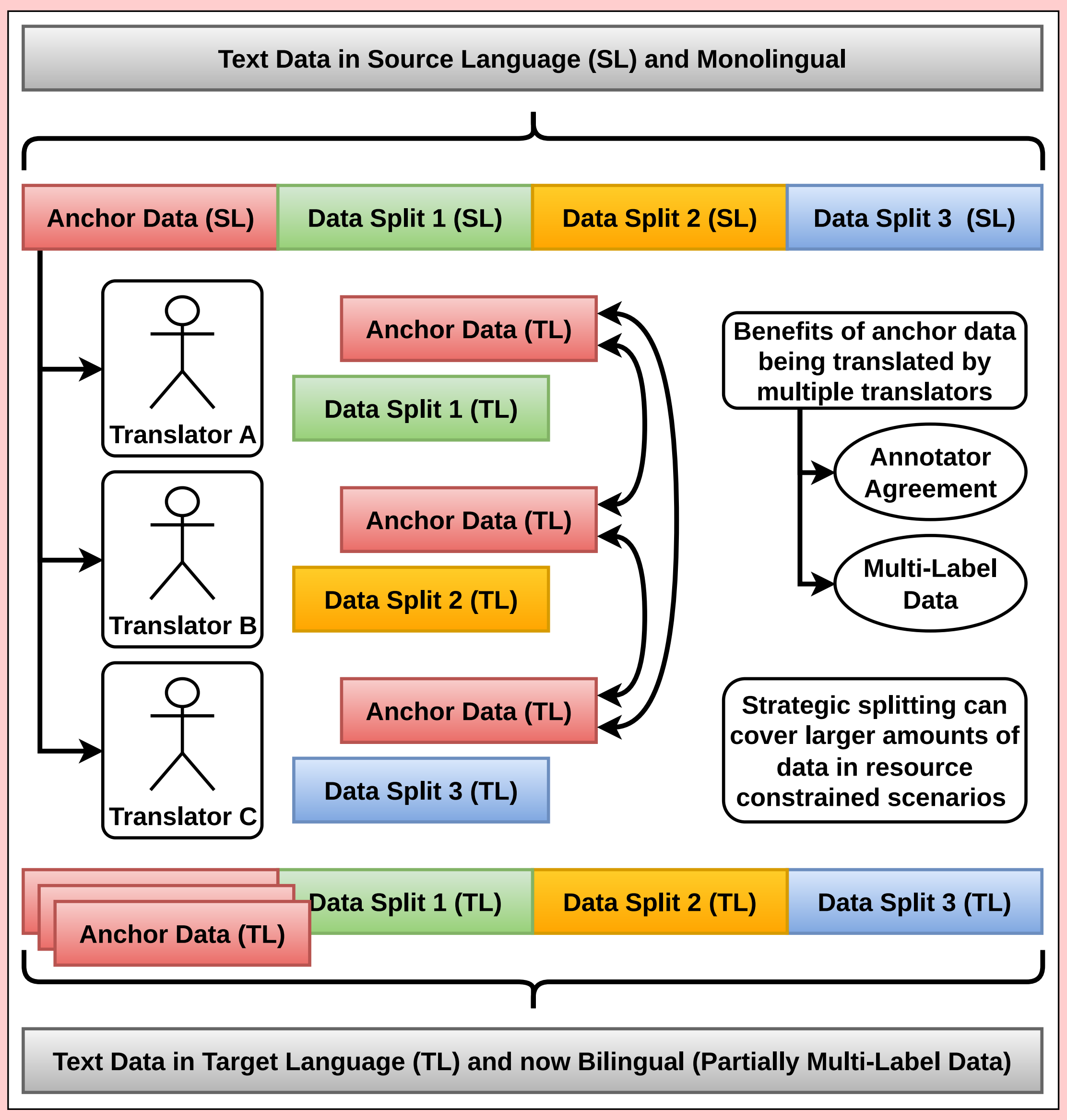
Toolkit in action for the **data acquisition**:

http://schuler-christian.de

Current state and future development found at:

https://github.com/christianschuler8989/CRAMT

Data Quality



Next Steps

Language data waiting to be collected.

References

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