TL-Explorer: A Digital Humanities Tool for Mapping and Analyzing Translated Literature

Alex Zhai[*]
EA 4073-GERiiCO
Univ. Lille, F-59000
Lille, France
alexhezhai@gmail.com

Zheng Zhang[*]
Schlumberger
Clamart, France
zzhang54@slb.com

Amel Fraisse
EA 4073-GERiiCO
Univ. Lille, F-59000
Lille, France
amel.fraisse@univ-lille.fr

Ronald Jenn
EA 4074-CECILLE
Univ. Lille, F-59000
Lille, France
ronald.jenn@univ-lille.fr

Shelley Fisher Fishkin
Stanford University
Stanford, CA
sfishkin@stanford.edu

Pierre Zweigenbaum
LIMSI-CNRS
Université Paris-Saclay
Orsay, France
pz@limsi.fr

Abstract

TL-Explorer is a digital humanities tool for mapping and analyzing translated literature, encompassing the World Map and the Translation Dashboard. The World Map displays collected literature of different languages, locations, and cultures and establishes the foundation for a variety of further analysis. It is comprised of three global maps for spatial and temporal interpretation. A further investigation into an individual node on the World Map — representing one edition or translation — leads to the Translation Dashboard. Collected translations are processed in order to build multilingual parallel corpora for a large number of under-resourced languages as well as to highlight the transnational circulation of knowledge.

Our first rendition of TL-Explorer was conducted on the well-traveled American novel, *Adventures of Huckleberry Finn*, by Mark Twain. The maps currently chronicle nearly 400 translations of this novel and the dashboard supports over 30 collected translations. However, the TL-Explore is easily extended to other works of literature and is not limited to type of texts, such as academic manuscripts or constitutional documents to name a few.

1 Introduction and Motivation

From a global perspective, human knowledge of culture and heritage has been shared, explored, and preserved for nearly centuries through translation. The art of translating texts is largely to thank for our ability to learn about and from other cultures, and vice versa. It is crucial to recognize that every person is shaped by their culture and identity. Hence, every body of knowledge, regardless of type of classification, is similarly impacted by specific historical, geopolitical, and sociocultural factors. TL-Explorer is created not only with this diversity in mind, but also as a tool to explore these nuances as they are reflected in translated literature.

TL-Explorer is designed to provide users with a feeling of continuity as they explore translated texts. The tool begins at a broad starting point — a global view of the entire collection of texts — and allows the user to smoothly zoom into a particular geographic region, individual editions or translations in that region, and specific chapters and paragraphs within the selected literature. The TL-Explorer uses a Geographic Information System (GIS) to create the World Map and NLP techniques to generate the Translation Dashboard.

2 Prior Work

2.1 Digital Humanities Mapping Tools

Hypercities (Presner et al., 2014) introduced a digital humanities mapping tool for exploring and interacting with the layered histories of city and global spaces. Spatialization tools or geographic information

[*] equal contribution

Proceedings of LaTeCH-CLfL 2020, pages 167–171
Barcelona, Spain (Online), December 12, 2020.

systems (GIS) such as Carto[1], Open Street Map (OpenStreetMap contributors, 2017), QGIS (QGIS Development Team, 2009), Harvard Worldmap (Guan et al., 2012), Spatial Data Explorer (Kollen, 2016) and Unfolding (Nagel et al., 2013) are also useful tools for digital humanities mapping.

2.2 Parallel Corpora Construction and Analysis Tools and Resources

There exist many construction and analysis tools for parallel corpora such as Uplug (Tiedemann, 2003), PENCIL (Kakoyianni-Doa et al., 2013) and The Sketch Engine (Kilgarriff et al., 2014), but there remain very little designed specifically for translated literature.

While the interpretive nature of literary translations has caused a lag in their adoption as a source for NLP development, multiple recent projects have developed parallel corpora based on well-known texts including the *Harry Potter* series and *Le Petit Prince*.

2.3 Alignment Visualization Tools

While there already exist alignment visualization tools such as ANNIS (Druskat et al., 2016), SWIFT Aligner (Gilmanov et al., 2014), Cario (Smith and Jahr, 2000), VisualTCA (Gomes et al., 2007) and MkAlign (Fleury and Zimina, 2007), most of them focus on word alignment. Further, even though some of these tools provide sentence alignment visualization, they are meant to be an intermediate step before the lexicon level. There are currently no other tools that allow users to explore data in a chapter-paragraph-sentence/word, coarse-to-fine fashion. Moreover, these tools are not oriented towards literary texts, which is more challenging for alignment approaches. Though alignment should be as confident as possible (Xu et al., 2015), this is complicated by the fact that a literary translation may include deliberate changes to the text inserted by the translator, and may not be a literal translation.

3 TL-Explorer

3.1 World Map Viewer

The World Map is the base tool in the TL-Explorer and provides a geographic display of collected translation information. It is separated into three maps: the Home Map, Heat Map, and Time Map, which display the same information in different views.

3.1.1 Home Map

The Home Map (see Figure 1) is the first map. It provides a global view of all the gathered texts, and each node represents one edition or translation. Nodes are placed at the location of publication, not based on the language of translation. Texts that are geographically close to each other are grouped into a cluster, represented by the light-yellow circles, with a number that reflects the number of texts in that cluster. A search function in the top left corner of the interface allows the user to search the map by title of the text. In the bottom left corner, a label identifies the number of texts being represented. The key informs the user of the types of languages represented on the map: English original, well-resourced language, medium-resourced language, and under-resourced language.

The World Map allows the user to zoom in from a global view to country view and even as close as specific streets. By clicking on a node, a pop-up displays the following information of the selected text: title, language, series/collection, edition, contributors (translators, editors, cover artists, and illustrators), date of publication, publisher, publisher city, and page count. If there is a digital version of the translation, the pop-up will include a link to it.

3.1.2 Time Map

The Time Map documents in chronological order developments of the literature of interest. The year is controlled by a scroll-bar below the map. As the years progress, nodes appear and accumulate. For example, Figure 1 displays the translation of *Adventures of Huckleberry Finn* up to the year 1940.

[1] https://carto.com/

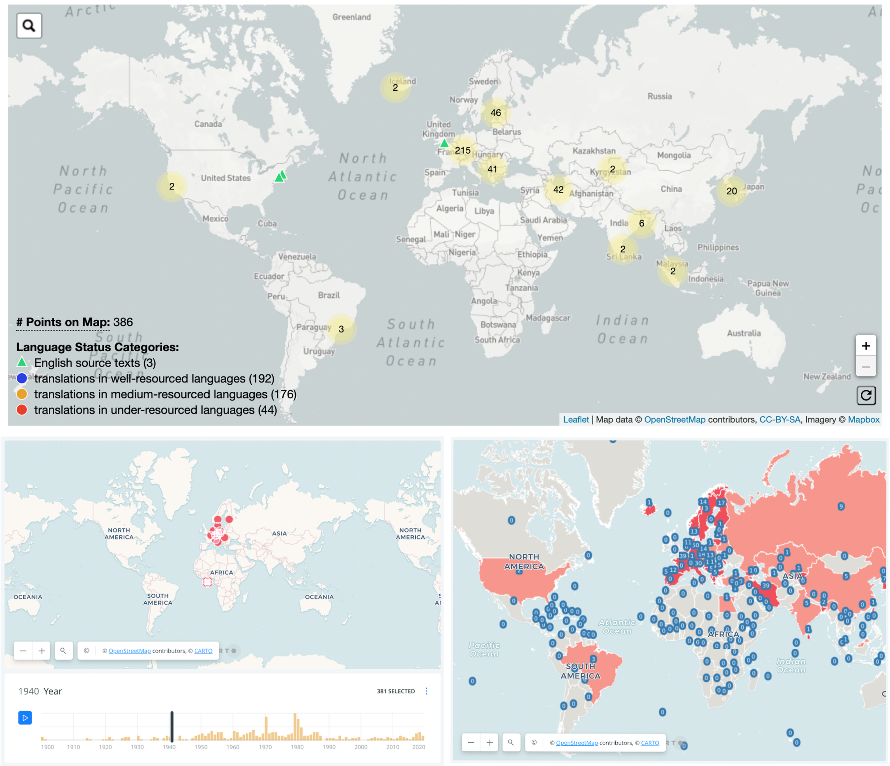

# Points on Map: 386
Language Status Categories:
English source texts (3)
translations in well-resourced languages (192)
translations in medium-resourced languages (176)
translations in under-resourced languages (44)
1940 Year
381 SELECTED

Figure 1: **Home Map** The width and height ratios of the maps have been changed to conserve space.

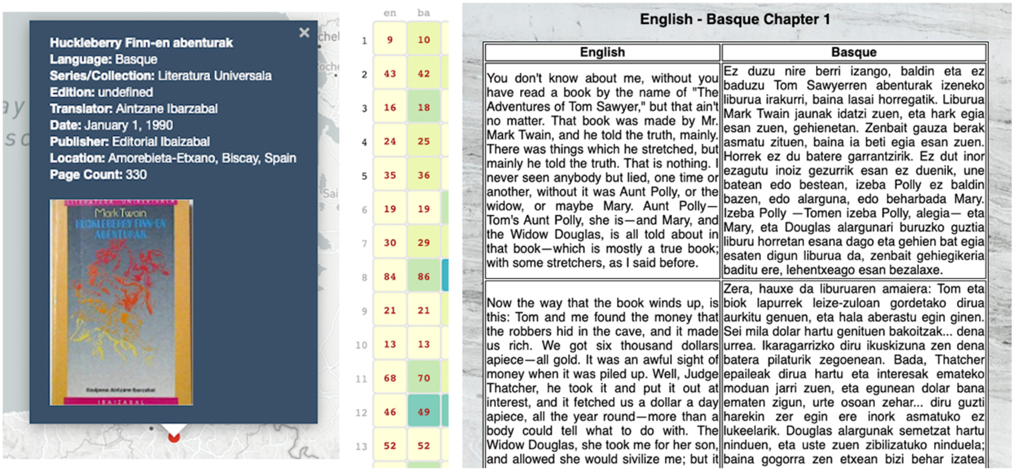
3.1.3 Heat Map

The third map is the Heat Map, which displays translation count for each country and provides insight on which areas are more frequently represented (see Figure 1). It is aggregated using the data at a certain time, and changes as more data is added, such as a new translation or language.

3.1.4 Crowdsourcing for Data Curation

There are currently 386 points or translations on the maps. However, the TL-Explorer is immensely scalable. The TL-Explorer has a form where the user can submit all the information they have pertaining to an individual translation. After data-vetting, the point will be added as a new node to the maps.

3.2 Translation Dashboard



Huckleberry Finn-en abenturak
Language: Basque
Series/Collection: Literatura Universala
Edition: undefined
Translator: Aintzane Ibarzabal
Date: January 1, 1990
Publisher: Editorial Ibaizabal
Location: Amorebieta-Etxano, Biscay, Spain
Page Count: 330

	en	ba
1	9	10
2	43	42
3	16	18
4	24	25
5	35	36
6	19	19
7	30	29
8	84	86
9	21	21
10	13	13
11	68	70
12	46	49
13	52	52

English - Basque Chapter 1

English	Basque
You don't know about me, without you have read a book by the name of "The Adventures of Tom Sawyer," but that ain't no matter. That book was made by Mr. Mark Twain, and he told the truth, mainly. There was things which he stretched, but mainly he told the truth. That is nothing. I never seen anybody but lied, one time or another, without it was Aunt Polly, or the widow, or maybe Mary. Aunt Polly—Tom's Aunt Polly, she is—and Mary, and the Widow Douglas, is all told about in that book—which is mostly a true book; with some stretchers, as I said before.	Ez duzu nire berri izango, baldin eta ez baduzu Tom Sawyerren abenturak izeneko liburua irakurri, baina lasai horregatik. Liburua Mark Twain jaunak idatzi zuen, eta hark egia esan zuen, gehienetan. Zenbait gauza berak asmatu zituen, baina ia beti egia esan zuen. Horrek ez du batere garrantzirik. Ez dut inor ezagutu inoiz gezurrik esan ez duenik, une batean edo bestean, izeba Polly ez baldin bazen, edo alarguna, edo beharbada Mary. Izeba Polly —Tomen izeba Polly, alegia— eta Mary, eta Douglas alargunari buruzko guztia liburu horretan esana dago eta gehien bat egia esaten digun liburua da, zenbait gehiegikeria baditu ere, lehentxeago esan bezalaxe.
Now the way that the book winds up, is this: Tom and me found the money that the robbers hid in the cave, and it made us rich. We got six thousand dollars apiece—all gold. It was an awful sight of money when it was piled up. Well, Judge Thatcher, he took it and put it out at interest, and it fetched us a dollar a day apiece, all the year round—more than a body could tell what to do with. The Widow Douglas, she took me for her son, and allowed she would sivilize me; but it	Zera, hauxe da liburuaren amaiera: Tom eta biok lapurrek leize-zuloan gordetako dirua aurkitu genuen, eta hala aberastu egin ginen. Sei mila dolar hartu genituen bakoitzak... dena urrea. Ikaragarrizko diru ikuskizuna zen dena batera pilaturik zegoenean. Bada, Thatcher epaileak dirua hartu eta interesak emateko moduan jarri zuen, eta egunean dolar bana ematen zigun, urte osoan zehar... diru guzti harekin zer egin ere inork asmatuko ez lukeelarik. Douglas alargunak semetzat hartu ninduen, eta uste zuen zibilizatuko ninduela; baina gogorra zen etxean bizi behar izatea

Figure 2: Translation Dashboard, example with a Basque translation

The field translation studies bridges comparative literature and corpus linguistics. While corpus linguists are likely to have at least basic programming skills and a broad familiarity with computational methods, the equivalent "digital humanities" training is less common in comparative literature. As a result, corpus linguists are more likely to focus on comparatively large-scale computational text analysis, and comparative literature scholars tend to conduct close examinations of a small number of texts.

The Translation Dashboard is designed as an adjunctive tool for researchers in translation studies grounded in the comparative literature tradition. It provides a reading environment that could display the visualizations and text in parallel in order to allow scholars to easily see patterns of structural divergence between the source text and translations at different levels of granularity.

Text is aligned at the paragraph, sentence, and world level using Natural Language Processing algorithms, including the IBM Models 1 and 2 for Statistical Machine Translation(Collins, 2011) and the Gale-Church Algorithm(Gale and Church, 1993).

3.2.1 Paragraph Count Analysis

After selecting a specific node on the World Map, the default view of the Translation Dashboard displays a per-chapter paragraph count, based on newlines and white space in that source text. The deviation in paragraph count between a source text and its translation is reflected in the color variation in the Heat Map within the table (see Figure 2, center image). An exceedingly high divergence from the source paragraph count alerts the scholar that there may be data cleaning issues (e.g. one instance where each line in a poem embedded in a narrative was treated as a new paragraph), but a moderate divergence can reflect the translator's deliberate stylistic choices about how the flow of the narrative should be rendered. A translation studies scholar in the literary tradition may use this information to select chapters for a close-reading analysis.

3.2.2 Paragraph Alignment

When the user selects an individual chapter in the text, they can view a display that presents both the original English chapter and the chapter in the translation. The tool displays the two parallel to each other to allow for easy comparison.

We divided chapters into 3 major categories based on the differences in their paragraph counts compared to the original English version: *exact-match*, *large-difference*, and *small-difference*. Different paragraph aligners may apply to different categories.

For *exact-match* chapters, our hypothesis is that their paragraphs were translated one to one. No further paragraph alignment methods are needed. This hypothesis has been confirmed for most of the *exact-match* cases by the human validation experiment.

Large-difference cases are normally caused by different ways of splitting quotations, so we provide a text pre-processing option before paragraph alignment when long quotations have been found under *large-difference* cases. This pre-processing option splits quotations into paragraphs according to the same standard in all translations. Experiments have shown that this action can significantly reduce differences in paragraph counts and sometimes move a chapter from the *large-difference* category to the *small-difference* category.

For the majority *small-difference* cases, we applied the Gale-Church algorithm(Gale and Church, 1993). Here we treat paragraphs as sentences so as to feed them into this sentence alignment algorithm. The tool is easy to use, and thus easy for a native speaker to provide feedback on the accuracy of the alignments.

4 Conclusion

Encompassing both of the World Map and the Translation Dashboard, the TL-Explorer allows for analysis of translated literature at an exceptional range of specificity. The World Map provides a global view that can be shrunken into exact coordinates and streets and the Translation Dashboard allows for intense analysis of two texts from entire works to specific sentences. The TL-Explorer similarly serves a purpose of preservation and globalization, representing a large number of under-resourced languages and a transnational circulation of knowledge.

References

Michael Collins. 2011. Statistical machine translation: Ibm models 1 and 2. *Columbia Columbia Univ*.

Stephan Druskat, Volker Gast, Thomas Krause, and Florian Zipser. 2016. corpus-tools. org: An interoperable generic software tool set for multi-layer linguistic corpora. In *LREC*.

Serge Fleury and Maria Zimina. 2007. Exploring translation corpora with mkalign. *Translation Journal*, 11(1).

William A Gale and Kenneth Church. 1993. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

Timur Gilmanov, Olga Scrivner, and Sandra Kübler. 2014. Swift aligner, a multifunctional tool for parallel corpora: Visualization, word alignment, and (morpho)-syntactic cross-language transfer. In *LREC*, pages 2913–2919.

Felipe Tassario Gomes, Thiago Alexandre Salgueiro Pardo, and Helena de Medeiros Caseli. 2007. Visualtca: Uma ferramenta visual on-line para alinhamento sentencial de textos paralelos. In *Anais do XXVII Congresso da Sociedade Brasileira de Computação-V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pages 1729–1732.

Weihe Wendy Guan, Peter K Bol, Benjamin G Lewis, Matthew Bertrand, Merrick Lex Berman, and Jeffrey C Blossom. 2012. Worldmap–a geospatial framework for collaborative research. *Annals of GIS*, 18(2):121–134.

Fryni Kakoyianni-Doa, Stefanos Antaris, and Eleni Tziafa. 2013. A free online parallel corpus construction tool for language teachers and learners. *Procedia-Social and Behavioral Sciences*, 95:535–541.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, pages 7–36.

Christine Kollen. 2016. Spatial data explorer: Providing discovery and access to geospatial data at the university of arizona.

Till Nagel, Joris Klerkx, Andrew Vande Moere, and Erik Duval. 2013. Unfolding–a library for interactive maps. In *International Conference on Human Factors in Computing and Informatics*, pages 497–513. Springer.

OpenStreetMap contributors. 2017. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`.

Todd Presner, David Shepard, and Yoh Kawano. 2014. *Hypercities thick mapping in the digital humanities*.

QGIS Development Team, 2009. *QGIS Geographic Information System*. Open Source Geospatial Foundation.

Noah A Smith and Michael E Jahr. 2000. Cairo: An alignment visualization tool. In *LREC*.

Jörg Tiedemann. 2003. *Recycling Translations – Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Ph.D. thesis, Uppsala University, Uppsala, Sweden. Anna Sågvall Hein, Åke Viberg (eds): Studia Linguistica Upsaliensia.

Yong Xu, Aurélien Max, and François Yvon. 2015. Sentence alignment for literary texts. *LiLT (Linguistic Issues in Language Technology)*, 12.