Preserving Endangered European Cultural Heritage and Languages Through Translated Literary Texts

Amel Fraisse[1], Zheng Zhang[2], Shelley Fisher Fishkin[3], Ronald Jenn[4]

[1] GERiiCO, Université de Lille
[2] LIMSI-CNRS, Université Paris-Saclay
[3] Department of English, Stanford University
[4] CECILLE, Université de Lille

Abstract

We present the interdisciplinary ROSETTA project which consists **in collecting all the translations worldwide of one fictional text** in order to build **multilingual parallel corpora for a large number of under-resourced languages**. Building such corpora is vital to help preserve and expand language and traditional knowledge diversity. These corpora will be useful to handle under-resourced languages in a number of interconnected research fields such as computational linguistics, translation studies and corpus linguistics. Our project taps into a wealth of translated versions of a single fictional text spanning a period of over a century. It consists in collecting, digitizing, transcribing and aligning translations of this text. Our data collection process is based on volunteer work from the scientific and scholarly communities, the power of the crowd and national libraries and archives. Our first experiment was conducted on the world-famous and well-traveled American novel "Adventures of Huckleberry Finn" by the American author Mark Twain. This paper reports on the parallel corpus that are now sentence aligned pairing English with Basque.

Keywords : under-resourced languages, parallel corpus, translated literary text

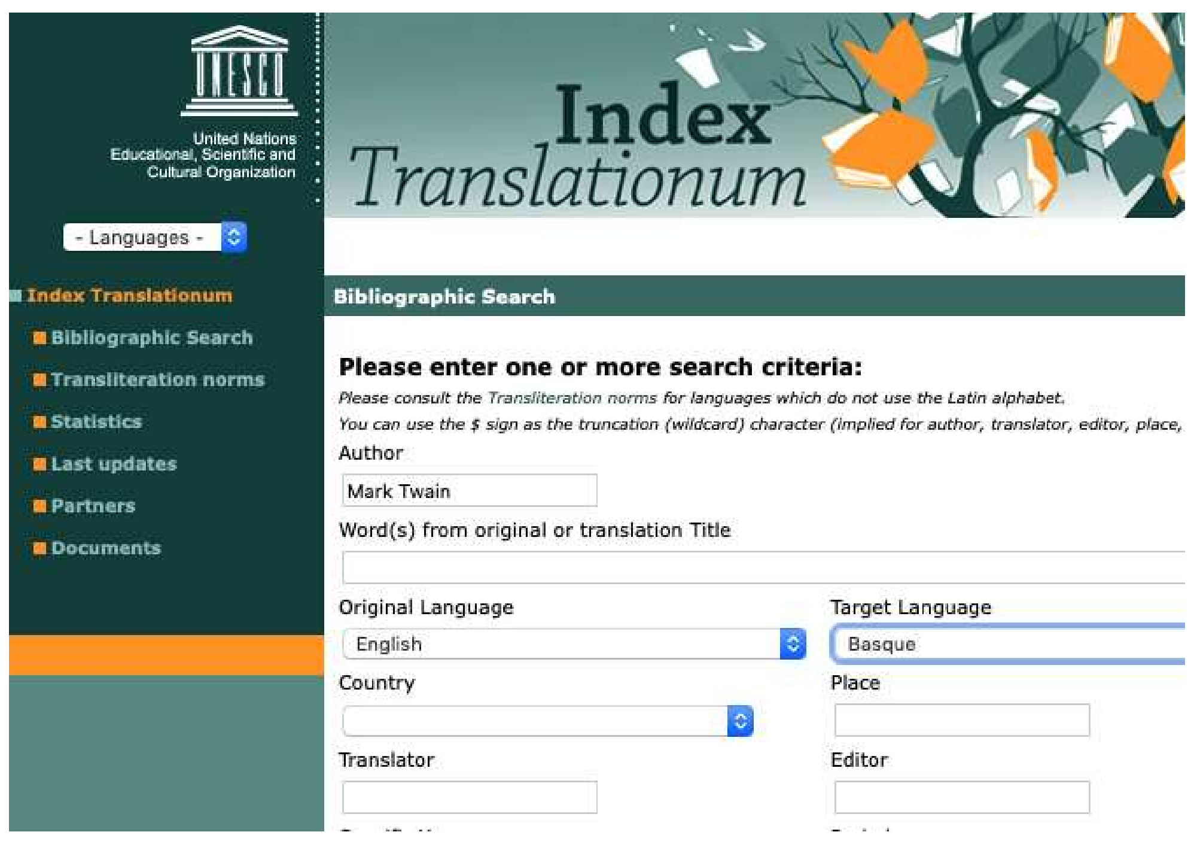
1 The example of Mark Twain's text for under-resourced languages

- Mark Twain's books are some of the most well-travelled texts on the planet.
- The advantage of using a work of fiction such as "Adventures of Huckleberry Finn", is that it uses a very broad vocabulary linked to every day life, which makes it a valuable asset for those languages that are currently lacking such computational resources.

Languages			
1. Afrikans	17. Estonian	33. Korean	49. Slovak
2. Albanian	18. Farsi	34. Latvian	50. Slovenian
3. Arabic	19. Finnish	35. Lithuanian	51. Spanish
4. Armenian	20. French	36. Macedonian	52. Swedish
5. Assamese	21. Georgian	37. Malay	53. Tamil
6. Basque	22. German	38. Malayalam	54. Tatar
7. Bengali	23. Greek	39. Marathi	55. Telugu
8. Bulgarian	24. Hebrew	40. Moldovan	56. Thai
9. Burmese	25. Hindi	41. Norwegian	57. Turkish
10. Catalan	26. Hungarian	42. Oriya	58. Turkmen
11. Chinese	27. Icelandic	43. Polish	59. Ukrainian
12. Chuvash	28. Indonesian	44. Portuguese	60. Uzbech
13. Croatian	29. Italian	45. Romanian	61. Vietnamese
14. Czech	30. Japanese	46. Russian	62. Yiddish
15. Danish	31. Kazakh	47. Serbian	
16. Dutch	32. Kirghiz	48. Sinhalese	

TABLE 1 – List of languages "Adventures of Huckleberry Finn" was translated into.

2 Data Collection



United Nations Educational, Scientific and Cultural Organization

Index Translationum

Bibliographic Search

Please enter one or more search criteria:

Please consult the Transliteration norms for languages which do not use the Latin alphabet.

You can use the $ sign as the truncation (wildcard) character (implied for author, translator, editor, place,

Author: Mark Twain

Word(s) from original or translation Title:

Original Language: English

Target Language: Basque

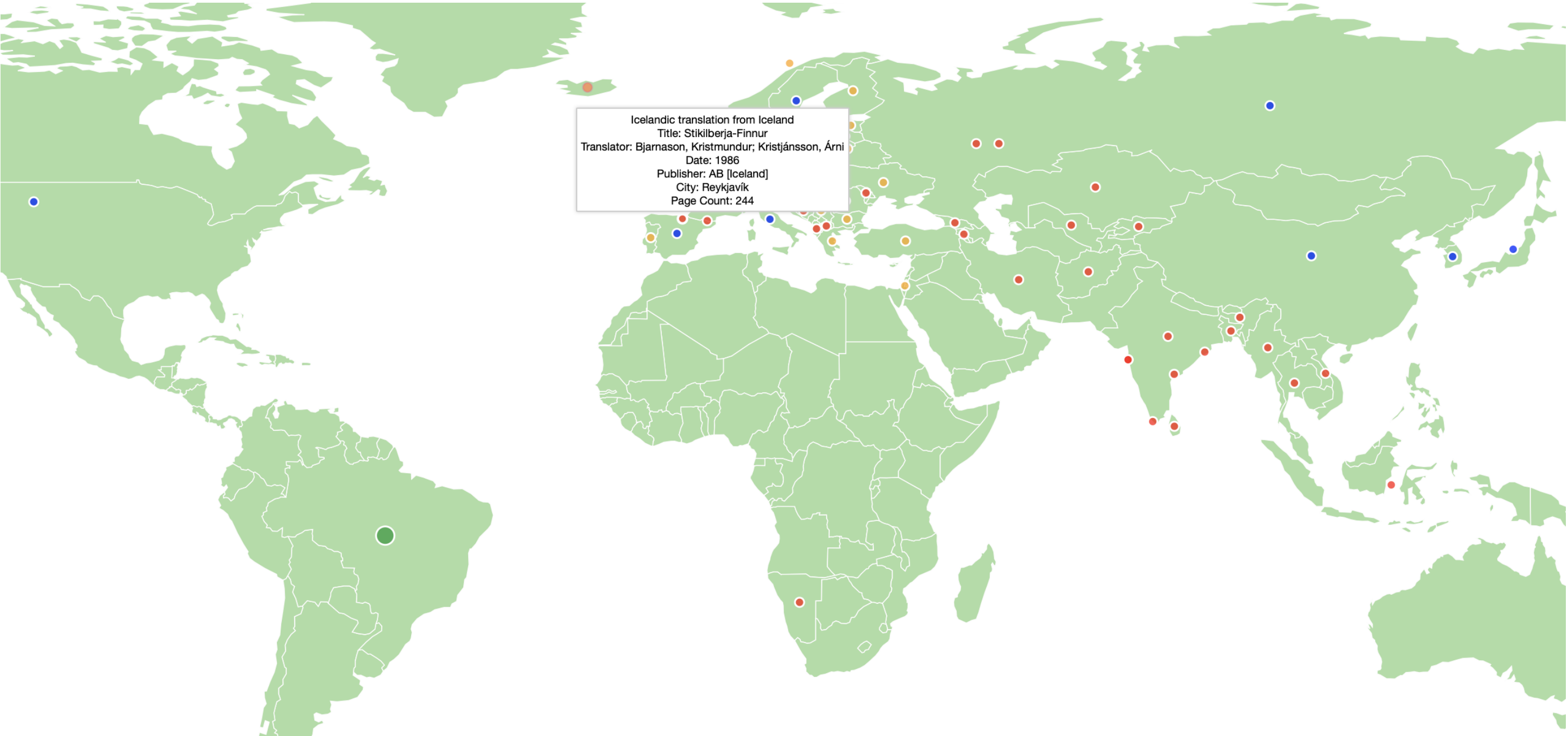
Country:

Place:

Translator:

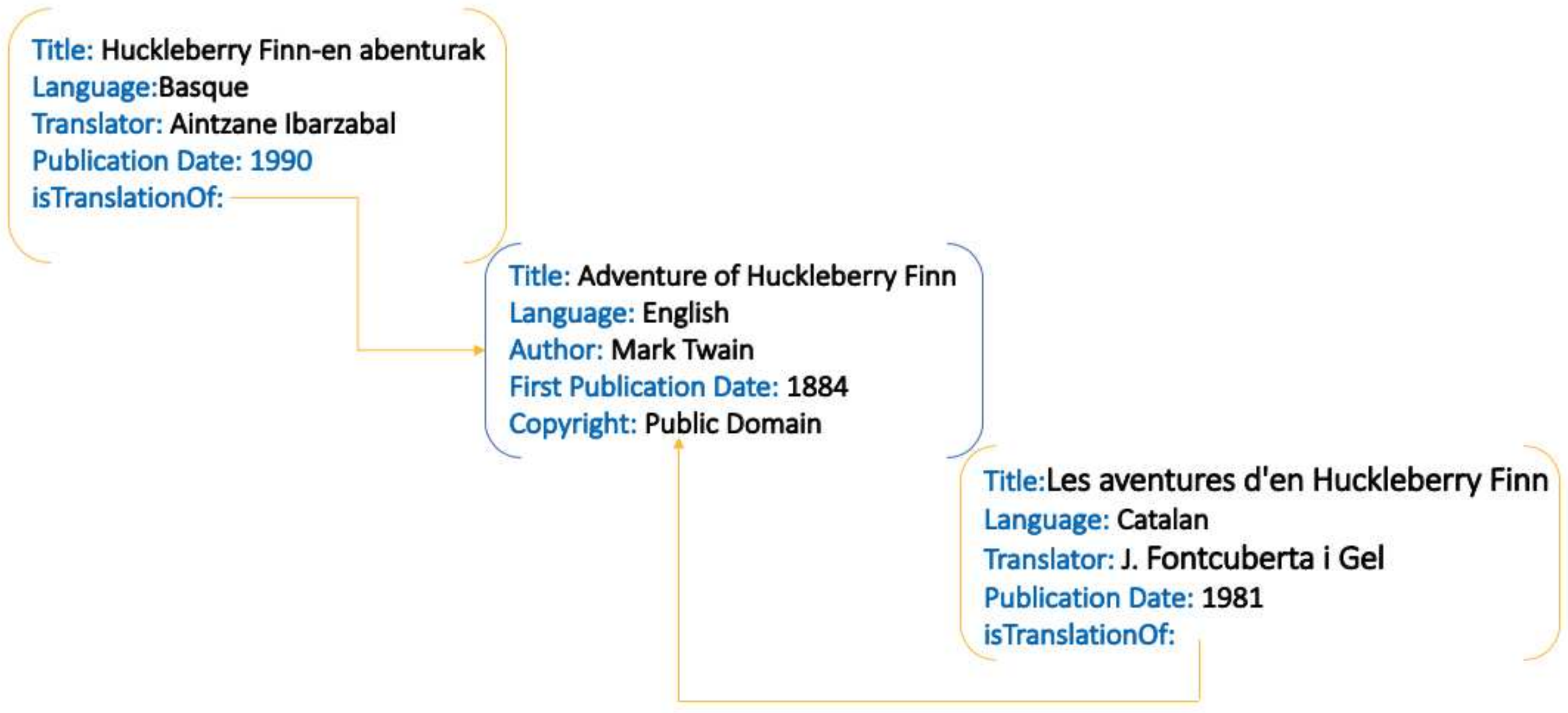
Editor:

FIGURE 1 – The UNESCO Index Translationum



Icelandic translation from Iceland
Title: Stikilberja-Finnur
Translator: Bjarnason, Kristmundur; Kristjánsson, Árni
Date: 1986
Publisher: AB [Iceland]
City: Reykjavík
Page Count: 244

FIGURE 2 – The global knowledge map representing existing translations of *Adventures of Huckleberry Finn*. In this map, the bubble over Iceland is highlighted, displaying the relevant information for the Icelandic translation.



Title: Huckleberry Finn-en abenturak
Language: Basque
Translator: Aintzane Ibarzabal
Publication Date: 1990
isTranslationOf:

Title: Adventure of Huckleberry Finn
Language: English
Author: Mark Twain
First Publication Date: 1884
Copyright: Public Domain

Title: Les aventures d'en Huckleberry Finn
Language: Catalan
Translator: J. Fontcuberta i Gel
Publication Date: 1981
isTranslationOf:

FIGURE 3 – The global knowledge diagram representing a subset of existing translations of Adventures of Huckleberry Finnin in Basque and Catalan.

1. Identifying existing translations : calling on the international community of Mark Twain scholars as well as Translation Studies scholars in order to identify existing translations in different languages.
2. Mining online digital libraries and national archives : using the title in the target languages, we crawled the web and mined online digital libraries and national archives in order to find the full texts.
3. Text transcription : We used the CrowdFlower platform to transcribe digital versions that came as images.

3 Experiments and Results

English - Basque Chapter 38	
English	Basque
Making them pens was a distressid-tough job, and so was the saw ; and Jim allowed the inscription was going to be the toughest of all. That's the one which the prisoner has to scrabble on the wall. But we had to have it ; Tom said we'd got to ; there warn't no case of a state prisoner not scrabbling his inscription to leave behind, and his coat of arms.	Idazteko lumak egitea oso lan nekagarria zen, eta zerra egitea ere bai, eta Jimek esan zuen inskripzioa egitea lanik gogorrena izango zela. Hori presoak horman egin behar zuen zirriborroa zen. Baina nahitanahiez izan behar genuen inskripzioa. Tomek esan zuen egin behar genuela ; ez zela estatuko preso bakar bat inskripzioa idatzita utzi ez zuenik, bere armarria eta guzti gainera.
"Why, Mars Tom, I hain't got no coat o' arms ; I hain't got nuffn but dishyer ole shirt, en you knows I got to keep de journal on dat."	— Baina, Tom jauna —esan zuen Jimek—, nik ez dut armarririk ; ez daukat alkandora zahar hau besterik, eta badakizu bertan egunkaria idatzi beharra dudala.
"Oh, you don't understand, Jim : a coat of arms is very different."	— Ez duzu ulertzen, Jim ; armarria bestelako gauza da.

TABLE 2 – Original text of the chapter 38 of "Adventures of Huckleberry Finn" and its translations in Basque

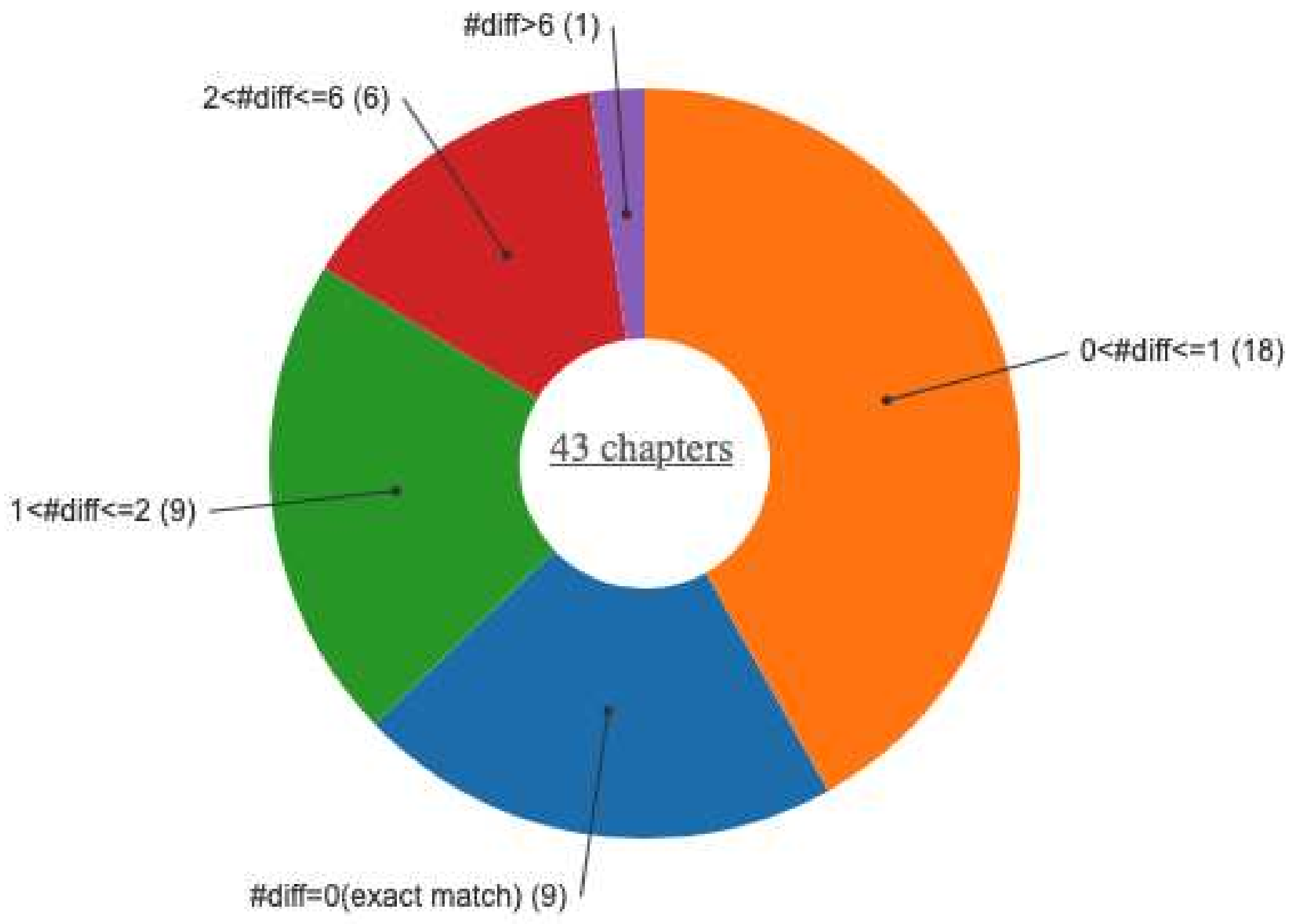


FIGURE 4 – Pie chart of the chapter distribution of Basque translation compared with the original English version.

Conclusions

- we proposed and experimented a new language source to build multilingual parallel corpora for a large number of under-resourced languages.
- It consists in collecting all the translations worldwide of one fictional text by means of collaboration between volunteers, researchers, scholars, digital libraries and especially national archives, which are in charge of storing valuable traditional knowledge for future use.
- This paper reports on one parallel corpus that are now sentence aligned pairing English with Basque.