HAL open science

A Sustainable and Open Access Knowledge Organization Model to Preserve Cultural Heritage and Language Diversity

Amel Fraisse, Zheng Zhang, Alex Zhai, Ronald Jenn, Shelley Fisher Fishkin, Pierre Zweigenbaum, Laurence Favier, Widad Mustafa El Hadi

To cite this version:

Amel Fraisse, Zheng Zhang, Alex Zhai, Ronald Jenn, Shelley Fisher Fishkin, et al.. A Sustainable and Open Access Knowledge Organization Model to Preserve Cultural Heritage and Language Diversity. Information, 2019, 10 (10), pp.303. 10.3390/info10100303 . hal-02565134

HAL Id: hal-02565134

https://hal.science/hal-02565134v1

Submitted on 6 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

information

MDPI

Article

A Sustainable and Open Access Knowledge Organization Model to Preserve Cultural Heritage and Language Diversity

Amel Fraisse [1,*], Zheng Zhang [2], Alex Zhai [2], Ronald Jenn [3], Shelley Fisher Fishkin [4], Pierre Zweigenbaum [2], Laurence Favier [1] and Widad Mustafa El Hadi [1]

1 Groupe d'Études et de Recherche Interdisciplinaire en Information et Communication (GERiiCO), Université de Lille, 59000 Lille, France; laurence.favier@univ-lille.fr (L.F.); widad.mustafa@univ-lille.fr (W.M.E.H.)
2 Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur-Centre National de la Recherche Scientifique (LIMSI-CNRS), Université Paris-Saclay, 91400 Orsay, France; zheng.zhang@limsi.fr (Z.Z.); Alex.Zhai@limsi.fr (A.Z.); pz@limsi.fr (P.Z.)
3 Centre d'Etudes en Civilisations, Langues et Littératures Etrangères (CECILLE), Université de Lille, 59000 Lille, France; ronald.jenn@univ-lille.fr
4 Department of English, Stanford University, 94305 California, CA, USA; sfishkin@stanford.edu
* Correspondence: amel.fraisse@univ-lille.fr

Received: 31 July 2019; Accepted: 24 September 2019; Published: 28 September 2019

Abstract: This paper proposes a new collaborative and inclusive model for Knowledge Organization Systems (KOS) for sustaining cultural heritage and language diversity. It is based on contributions of end-users as well as scientific and scholarly communities from across borders, languages, nations, continents, and disciplines. It consists in collecting knowledge about all worldwide translations of one original work and sharing that data through a digital and interactive global knowledge map. 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. Building such corpora is vital in preserving and expanding linguistic and traditional diversity. 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 10 parallel corpora that are now sentence-aligned pairs of English with Basque (an European under-resourced language), Bulgarian, Dutch, Finnish, German, Hungarian, Polish, Portuguese, Russian, and Ukrainian, processed out of 30 collected translations.

Keywords: library and information science; knowledge organization systems; information flow; knowledge diversity; under-resourced languages; parallel corpora; translated texts

1. Introduction

The impact of the digital revolution on the preservation, organization, and sharing of human knowledge encoded by languages constitutes an extraordinarily rich phenomenon, characterized by both productive opportunities and obstacles and threats.

In the first place, digital has created tremendous opportunities in terms of accessing knowledge. Indeed, more people and especially persons belonging to minority communities can enjoy knowledge more easily, quickly and cheaply. New technologies also constitute a step forward in terms of public inclusion and awareness. In fact, the general public can be included and integrated, thanks to social networks and collaborative platforms, to provide mass dissemination of human knowledge.

Nevertheless, there are numerous barriers that prevent the sustaining of knowledge diversity. Language is the most important barrier; as language diversity is decreasing, the preservation and

transmission of such knowledge is at risk. The ever growing scientific and political interests in making knowledge open, accessible and sustainable has sparked major interest in many parts of the scientific community. Some disciplines have been concerned with problems of knowledge preservation, organization and dissemination for a long time. Library and Information Science (LIS) is such a discipline [1]. As a gateway to knowledge and culture, the field of LIS holds a long history on collecting, storing, organizing, and sharing access to knowledge. To this purpose, Knowledge Organization Systems (KOS), Information Retrieval Systems (IRS) and metadata exchange standards, among others, have been developed to meet the opportunities arising through the development of new technologies. Collections of the world's great libraries have been made available to the public through large-scale digitization. The Online Computer Library Center (OCLC) dedicated to the public purposes of furthering access to the world's information produces and maintains WorldCat, the largest online public access catalog (OPAC) in the world. WorldCat itemizes the collections of 72,000 libraries in 170 countries and territories. Multilingual online digital libraries and archival projects collect documents and make them available to a wide audience: the Wikisource project (https://wikisource.org), an online digital library of free content textual sources, the Internet Archive project (https://archive.org) building a digital library of Internet sites and other cultural artifacts in digital form such as books and audio records, or the Gutenberg project (https://www.gutenberg.org) offering over 56,000 free written and audio eBooks and especially older works for which copyright has expired in more than 50 under-resourced languages. Those ongoing projects have made and continue to make significant progress in the preservation of knowledge and language diversity.

A large but still modest number of languages, close to a hundred, have the so-called Basic Language Resource Kit (BLARK): monolingual and bilingual corpora, machine-readable dictionaries, thesauri, part-of-speech taggers, morphological analyzers, parsers, etc. [2,3]. This means that, as mentioned by Scannell [4], over 98% of world languages lack most, and usually all, of these language resources. Consequently, these languages and subsequently, knowledge encoded in these languages are threatened and their preservation is at risk. Digital language resources can help prevent the disappearance of diverse knowledge systems, ensure their preservation and transmission, and foster their cross-fertilization.

Even for well-endowed languages, parallel corpora, a valuable resource for sustaining linguistic diversity, are rare despite the great need there is for them. Such corpora are often used for testing new tools and methods to develop under-resourced languages. Because translated texts are de facto parallel corpora, and because we know for a fact that translated language materials do exist in under-resourced languages, using these translations can help build cheap, efficient, and reliable corpora for the purpose of preserving under-resourced languages. Those translations are mostly available in print and still awaiting digitization. They are all the more precious because, when translation does occur, it is currently into commercially dominant languages [5–9].

2. The Role of Library and Information Science in Building a Global, Shared Knowledge Community

More than a century ago, Paul Otlet, the pioneer of Documentation Studies (known today as Library and Information Science or LIS), envisioned a universal compilation of knowledge and the technology to make it globally available. He wrote numerous essays on how to collect and organize the world's knowledge, culminating in two books [1,10]. As described in [11],

> for Otlet the main questions were: how best was order to be introduced into this proliferating, disorderly mass in such a way that progress in the world of learning could continue efficiently and effectively ? How could rapid developments in all areas of knowledge, so characteristic of the modern period, be mobilised for the benefit of society ? How could the international flow of information, then obstructed (as it still is) by political, social and linguistic barriers on the one hand, and by cumbersome, unresponsive systems of publication, distribution and bibliographic processing on the other, become more open and more effective? How could accurate, up-to-date, 'integrated' information tailored specifically and exactly to particular

needs be derived from this mass and reworked to a form ensuring immediate and optimal usefulness. How could this especially processed information be made available without hindrance or delay, whatever such potentially infinite, unpredictable needs might be?

One of the most active vectors has been the emergence of digital tools as a new dissemination model for knowledge in a global context. The ever growing number of digital documents and scientific and political interests in making them openly available all over the world has led to the creation of new digital collections in a broad range of fields and languages. Several Registries of Open Access Repositories (ROAR) hosted by national and international Organization or universities have been developed. For example, The Library of Congress's American Memory collection (https://www.loc.gov) features approximately 164 million items in virtually all formats, languages, subjects, and periods. These collections are broad in scope, including research materials in more than 470 languages and multiple media. The Europeana collection (https://www.europeana.eu/), launched in 2008 and funded by the European Commission, contains over fifteen million digitized paintings, drawings, maps, photos, books, newspapers, letters, diaries, films, newsreels, etc., from fifteen hundred institutions.

However, the language barrier is a key issue that Knowledge Organization Systems (KOS) have to address. Indeed, over time, the gap between languages of dominant nations or civilizations and other languages has been growing. Although KOS include knowledge encoded in under-resourced languages, their use and exploration is still limited. Thus, in an increasingly globalized context, multilingualism has become a major preoccupation for the field of LIS and in particular for Knowledge Organization Systems which have to be as fair as possible [12–14] to ensure and sustain knowledge diversity.

3. Related Work

Over the last few years, there has been a growing interest and awareness among the scientific community and locally among advocates of minority languages in sustaining and expanding the existing resources in under-resourced languages and digitizing them in order to preserve and promote knowledge and language diversity. Sustaining knowledge encoded in under-resourced languages first needs the development of digital language resources. Parallel corpora are one example of such resources. Building such corpora to develop under-resourced languages is becoming the focus of many Natural Language Processing (NLP) scientific groups. Unlike monolingual corpora, the number of available multilingual parallel corpora is limited. Building these corpora presupposes the existence of translated language materials in under-resourced languages, where such resources are mostly available in print and are awaiting digitization. Some research focused on religious texts such as the Bible as a relevant source to compile massively parallel corpora [15]. This line of research, which entails the compilation of many parallel corpora, has broken new ground and allowed computational linguistics to handle an important number of under-resourced languages. More recently a Bible corpus was created based on freely available resources with over 900 translations in over 830 language varieties [16]. In [17], the authors built a massively parallel corpus based on 100 translations of the Bible, emphasizing difficulties in acquiring and processing the raw material. In [4], web-crawled corpora for many minority and under-resourced languages have been created and several open NLP tools have been developed for these languages in collaboration with native speakers. In [18,19], the authors created parallel aligned POS tagged corpora in 12 major Indian languages (including English) with Hindi as the source language in the domains of health and tourism.

For European languages, there is the JRC-Acquis parallel corpus [20], the first of the sentence-aligned and pre-processed corpora distributed by the European Commission. In its latest version, it comprised 22 languages, that is to say all of the current 24 official EU languages except for Irish and Croatian.

There are also parallel corpora related to translated literary works (e.g., "Harry Potter", "Le Petit Prince", and "Master i Margarita") or translations from the web, mostly available for a

set of closely related languages [16,21]. Most of these texts mainly concern well-endowed largely known languages.

While there already exist alignment visualization tools of parallel corpora, such as ANNIS [22], SWIFT Aligner [23], Cario [24], VisualTCA [25] and MkAlign [26], most of them focus on word alignment. Even if some of these tools provide sentence alignment visualization, they just serve as an intermediate step before lexicon level. There is no existing knowledge organization system allowing users to explore knowledge from coarse to fine analyses. Moreover, sustaining knowledge diversity is also the main focus of other research fields that are interested in how knowledge is produced, translated, adapted, circulated and received by local cultures. In [27], the authors examined how one data hub is working to become a relevant and useful source in the Web of big data and cloud computing. They focused on the OCLC's WorldCat data and explained how OCLC has begun work on the knowledge graph and its active involvement with *Schema.org* to make knowledge useful throughout the Web. In a recent work [28], the authors collected, analyzed and visualized through an online platform (https://routes-traductions.huma-num.fr) metadata of nine masterpieces considered as sources of knowledge and technology, from philosophy and mathematics to geography and medicine. One such author, *Hergé*, has had his books translated into many languages. In [29], the author introduced and described a new model for data curation and sharing by inviting colleagues around the world to collaborate on Digital Palimpsest Mapping Projects (DPMPs), or "Deep Maps". "Deep Maps", curated collaboratively by scholars in multiple locations, would put multilingual digital archives around the globe in conversation with one another, using maps as the gateway.

4. New Paradigm for a Sustainable Knowledge Organization Model

We propose a new paradigm that will permit different types of volunteers and contributors to take a part in the knowledge organization process in an efficient and dynamic way: while using the knowledge system, volunteers and contributors who know the local culture and language can participate by adding missed knowledge about a given translation of an original work. Volunteers and contributors could be scholars or simply citizens interested in preserving knowledge diversity.

We define global knowledge *GK* about an original work *ow* as a set of knowledge *k* about different translations of *ow*:

$$\mathrm{GK_{ow}} = \{K_{ow}^{t_1}, K_{ow}^{t_2}, \ldots, K_{ow}^{t_n}\} \tag{1}$$

where a knowledge $K_{ow}^{t_i}$ about a given translation t_i is a set of key properties as:

$$\mathrm{K_{ow}^{t_i}} = \{title, target_language, source_language, translator_name, publication_date, editor, full_text, comments\} \tag{2}$$

Figure 1 shows our model of representing some translations of *Adventures of Huckleberry Finn*, along with a set key properties. Each translation has a relationship *isTranslationOfWork* to the original work or to another translation from which it was translated. There can be multiple translations into the same language.

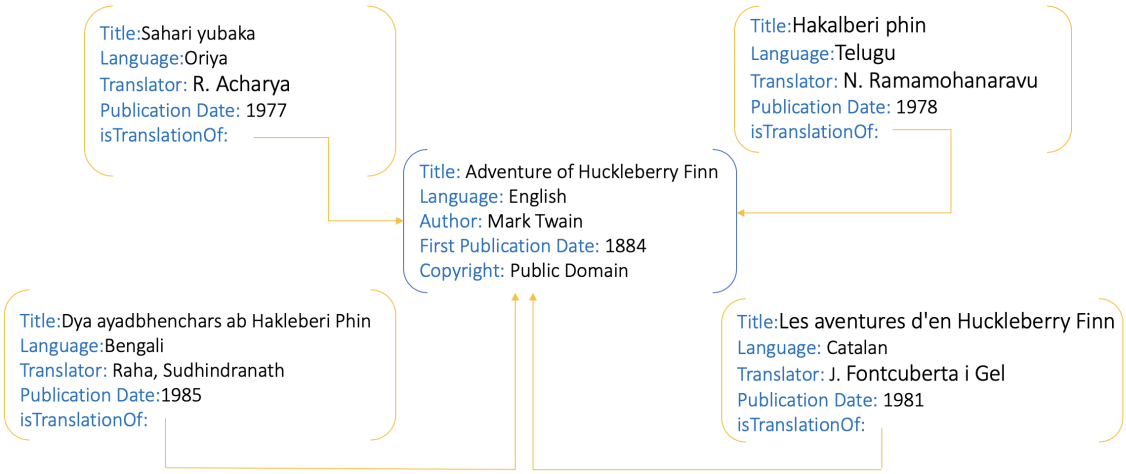

Figure 1. The global knowledge diagram representing a subset of existing translations of *Adventures of Huckleberry Finn* in four languages.

4.1. Why Focus on Translations?

The World's cultural and knowledge heritage is shared by being translated—it is how we learn about other cultures and how other cultures learn about us. Our model is created with the aim to recognize the fact that every body of knowledge is impacted locally as it is both written and read in a specific context and culture. According to the UNESCO *Index Translationum* (http://www.unesco.org/xtrans/) Figure 2 shows the world's twenty most translated authors. The mystery writer Agatha Christie has the distinction of being the world's most-translated author with 7233 translations—almost 3000 more than the next most popular, Jules Verne.

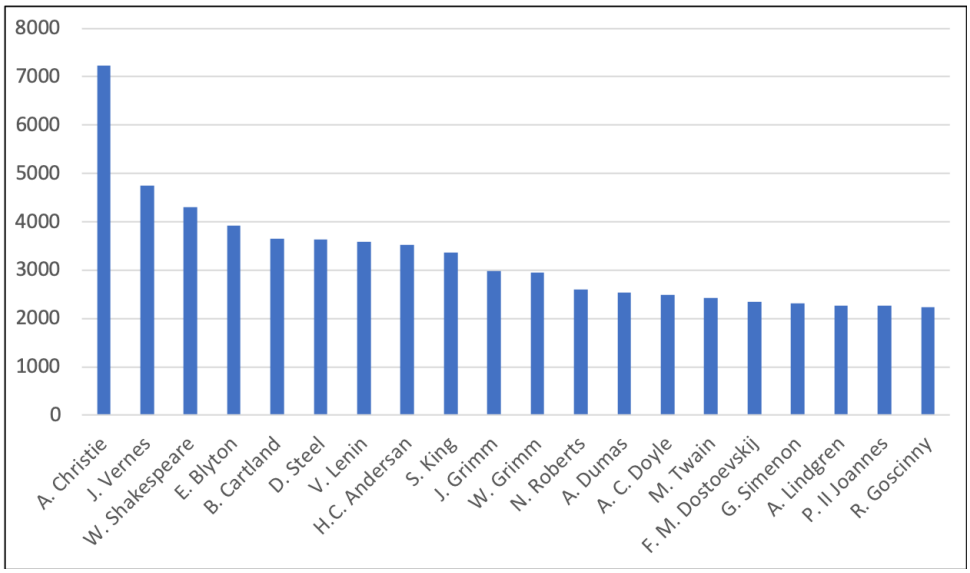

Figure 2. The world's twenty most translated authors.

4.2. Static Versus Interactive Knowledge Sharing Process

Unlike existing knowledge sharing models used by most digital libraries and collections, we propose a new interactive model allowing end-users and volunteer scholars to contribute and to share their knowledge about an original work through an interactive and online global knowledge map (Figure 3) (called Deep Maps in [29]). The global knowledge map displays all knowledge about all existing translations of a given original work. Each translation is represented by a node on the

world map, which could be considered as "completed" when all required knowledge is provided and "partially completed" when it lacks some knowledge such as the translator name or the full text. Nodes are updated incrementally by end-users and scholars through the map.

Figure 3. 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.

4.3. Non-Collaborative and Exclusive Versus Collaborative and Inclusive Knowledge Curation Processes

As described above, digital libraries projects have taken up the role of data curation facing a range of highly challenging issues considering the diversity of knowledge encoded in different languages and in particular those encoded in under-resourced ones. Unlike the existing curation model where knowledge is collected only by professional librarians or researchers, our model is based on a new paradigm that includes transnational volunteers in the data curation process. Indeed, during the map exploration, the end-user could edit any node to add missing knowledge. Content added by users are then evaluated and validated by experts before being shared.

5. Experiments and Results

We conducted a first experiment of our model on Mark Twain texts. Mark Twain's books are some of the most well-travelled texts on the planet. As the UNESCO Index Translationum (http://www.unesco.org/xtrans/) shows (Figure 2), the American writer is ranked 15th in the Top-20 of the most translated authors worldwide. His works have been translated into many languages [30] including under-resourced languages. The novel *Adventures of Huckleberry Finn* [31] is one of the most commonly translated of his books. Table 1 shows the scores of languages into which the book has been translated. The list includes 22 under-resourced languages: Assamese, Basque, Bengali, Burmese, Catalan, Chuvash, Farsi, Hindi, Indonesian, Kazakh, Kirghiz, Malayalam, Marathi, Moldovan, Oriya, Sinhalese, Tamil, Tatar, Telugu, Turkmen, Uzbek and Yiddish. In many of these languages, there have been multiple translations over time, reflecting different moments in history, and different ideological perspectives on the part of the translators or publishers, as well as different attitudes towards the US, childhood, minorities and minority dialects, race and racism, etc. Usually parallel corpora focus on very specific and specialized domains which can be efficient but also show limitations for machine translation. 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.

Table 1. List of languages *Adventures of Huckleberry Finn* was translated into.

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	

5.1. Data Curation

We started out by calling on the international community of Mark Twain scholars as well as Translation Studies scholars in order to identify existing translations in different languages. A globalized and transnational approach to Mark Twain is currently trending within that community. There is a growing interest in how Mark Twain's ideas and texts were translated and interpreted in different languages and especially the rarer ones.

In addition to the bibliographical survey carried out in [30], the Twain scholars and volunteers community provided us with a compiled list of additional references through, for example, field research at the UNESCO in Paris. In the compiled list resulting from those different inputs, each item includes the title in the target language, the first year of publication, the name of the translator and the publisher, when available. 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. In some cases, we came across the full online version that was in the public domain (provided by public institutions) in which case we downloaded them, whatever their format. When dealing with versions in pdf or epub format, we converted them into text format that could later be processed. There were other instances when we knew of an existing version but it was not readily available online. In that case, we turned to the national libraries and archives and asked them if they were willing to collaborate with us by digitizing their printed versions. Table 2 shows collected metadata (the title, the language, the translator name, the year of publication, and the publisher house) as well as full text files concerning *Adventures of Huckleberry Finn* translations. In total, we collected 62 metadata and 30 full text files concerning 22 under-resourced languages. Volunteer contributors and scholars provided us with 34 metadata and 7 full-text files. The crowdsourcing provided us with 18 metadata and full text files, and five translations were collected by crawling different digital libraries collections.

Table 2. Collected knowledge about the *Adventures of Huckleberry Finn* translations.

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

5.2. A Crowdsourcing Approach for Text Collection and Transcription

Due to the significant number of existing translations and the growing number of digital versions made available online, the crowdsourcing allowed us to gather data that would have otherwise been beyond our reach. Crowdsourcing helped reduce the amount of time spent on the task, increase the

variety and the range of the data covered (such as identifying translations which are not indexed in public databases). We used the figure-eight (https://www.figure-eight.com/) crowdsourcing platform. The parameterization of the experiment was as follows: as we are looking for translations over the world, we have not limited the geographic location of the contributors. Each task consisted of a set of nine questions (i.e., units in the figure-eight terminology) and completing the task will earn $0.25 (instead of $0.15 recommended by figure-eight). First, we asked people to use search engines or online catalogs to look for existing translations in their native language. Then, we asked them if they could find the translator's name, the first year of publication, the publishing house, the URL of the cover, the bibliographic record, and available public digital versions. Because of the complexity of the task, the crowdsourcing approach did not appear to be the best option. We assumed that the cultural background of crowdsourcing workers would not allow them to complete the task efficiently but it turned out that they managed to provide us with valuable and reliable information. One week after launching the job on the figure-eight platform, we received 710 judgments covering 31 different languages. After data cleaning, we collected 29 translations in 22 languages of different formats (html, text, pdf, and epub).

In a second step, we used the figure-eight platform to transcribe digital versions that came as images, whether from local institutions or collected from the web. The task asked workers to transcribe the text of one page as is, including misspellings and abbreviations. Figure 4 shows the example of the transcription task for the Bengali text.

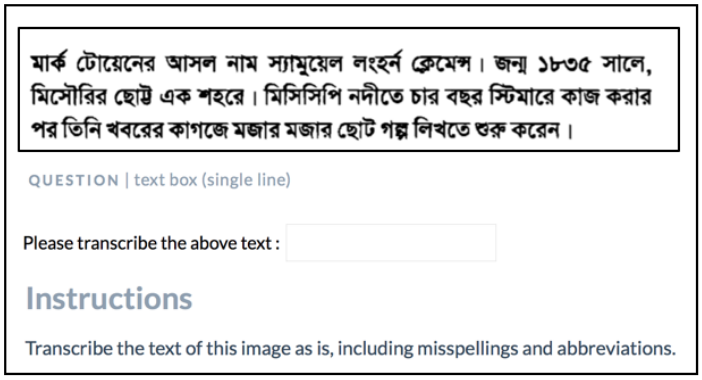

Figure 4. Example of the Bengali transcription task. The digitized image is at the top. Below it is the task instruction.

5.3. Data Alignment for Building Parallel Corpora

In a previous work [32], we collected a corpus containing *Adventures of Huckleberry Finn* translations in 22 different languages as a basis for developing parallel corpora for under-resourced languages. In this study, we focused on the Basque translation as it is an under-resourced language. The paragraph and word alignment algorithms are modest compared to the current cutting-edge of NLP, but the tool is nonetheless significant as an example of how to apply "just enough" NLP in the service of research priorities shaped by another field such us Library and Information Science or Translation Studies.

To get paragraph alignments in each chapter, we divided chapters into three 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. Thus, 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 it 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 started with applying the frequently used Gale–Church Aligner [33]. Here, we treated paragraphs as sentences so as to feed them into sentence aligners.

We applied IBM word-alignment models (of which there are five iterations) [34] to our aligned paragraphs. For each word in the original English corpus, the IBM models produce a list of possible translations in the target language and calculate a probability score for each English–Translation pairing.

To assess the reliability of these alignment pairs, we used a Top-5 accuracy score as our evaluation metric. First, we selected the most frequent words from the original English text. We then extracted the five translated words in the target language with the highest probability scores, produced by the aforementioned IBM models. Next, we fed these five translations into a machine translation program (Google Translate) in the target language and checked for an appearance of the original English word in the resulting English translation.

With the goal of producing high accuracy word alignment pairs, we executed the process above with varying parameters—different IBM Models (Models 1 and 2) and pre-processing setups (i.e., removing punctuation marks, lower-casing tokens, and applying different tokenizers). After these evaluations, we found that applying the most simple algorithm, IBM Model 1, to the text without any of the aforementioned pre-processing steps produced the most accurate results, with an acceptable Top-5 accuracy score of 40.0% for the 50 most frequently-appearing English words.

5.4. The Rosetta Dashboard for Fine-grained Knowledge Circulation Analysis

In addition to the online global knowledge map, we developed an online platform, the *Rosetta Dashboard*, allowing scholars in different research fields to access knowledge through a parallel reading environment and visualizations. The *Rosetta Dashboard* allows users to easily see patterns of structural and cultural divergence between the source text and translations, at different levels of granularity. In this section, we show an example for the user side of it. Starting with the main global knowledge map page, as shown in Figure 5, each cell value corresponds to the number of paragraphs in its chapter (row) and language version (column). A cell's color varies with the difference in the paragraph count, compared with the original English version. By referring to the legend below, we can have immediate general impressions of the paragraph count, such as: the paragraph count for the Basque and Hungarian translations are closer to the original English version compared to other languages (columns), and the number of paragraphs of Chapter 1 and Chapter 10 are close for all ten translations.

By clicking any target language label at the top of the heat map, a user can jump to the corresponding translation page, as shown in Figure 6. On this page, all 43 chapters are categorized into several groups by their paragraph count difference. For instance, in the Basque translation, we see that there are 36 chapters whose paragraph count difference is equal to or less than 2, including nine chapters where the paragraph count is an exact match to the original version.

We may apply different prepossessing and paragraph alignment algorithms to different kinds of chapters. This pie chart also provides useful information for algorithms selection.

A user can jump back to the main page by clicking the "43 chapters" label in center of the pie chart. Selecting the chapter cell in the target language (column) takes the user to the paragraph alignment page for the corresponding language and chapter, as shown in Figure 7. The paragraph alignment page contains re-organized paragraphs from two languages. The aligned ones are shown close to each other in horizontal position. Mousing over one paragraph will highlight the related aligned paragraphs.

Figure 5. Paragraph counts from Chapter 1 to Chapter 10 in English, Basque, Bulgarian, Dutch, Finnish, German, Hungarian, Polish, Portuguese, Russian and Ukrainian of the novel *Adventures of Huckleberry Finn*. Heat map color varies with the difference of paragraph counts in the corresponding chapter.

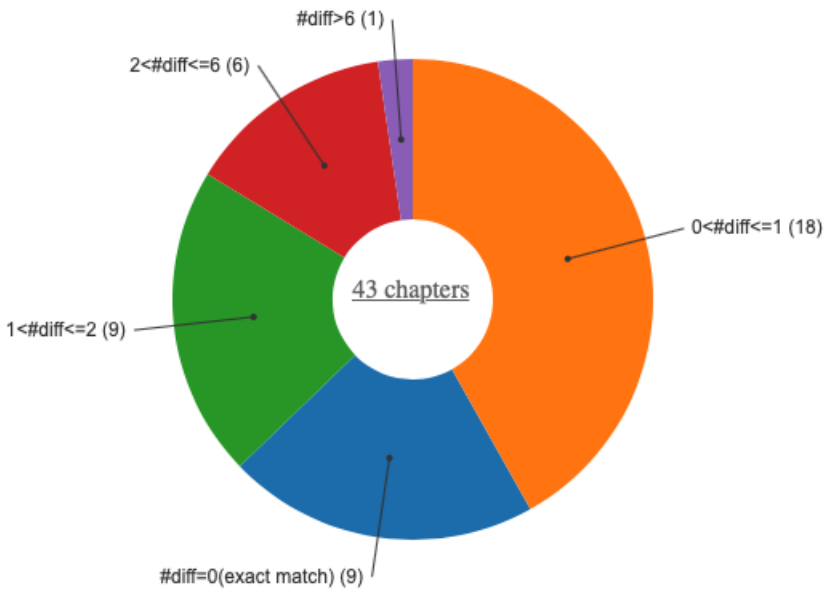

Figure 6. Pie chart of the chapter distribution of Basque translation compared with the original English version.

English - Basque Chapter 36

English	Basque
soon as we reckoned everybody was asleep, that night, we went down the lightning-rod, and shut ourselves up in the lean-to, and got out our pile of fox-fire, and went to work. We cleared everything out of the way, about four or five foot along the middle of the bottom log. Tom said he was right behind Jim's bed now, and we'd dig in under it, and when we got through there couldn't nobody in the cabin ever know there was any hole there, because Jim's counterpin hung down most to the ground, and you'd have to raise it up and look under to see the hole. So we dug and dug, with the case-knives, till most midnight; and then we was dog-tired, and our hands was blistered, and yet you couldn't see we'd done anything, hardly. At last I says:	Gau horretan, denak lotan izango zirela iruditu zitzaigunean, tximist orratzaren burdinan behera jaitsi ginen, estalpe barruan sartu, atea itxi, eta handik egur argitsua atera genuen eta lanera joan ginen. Metro t'erdian edo zeuden gauza guztiak kendu genituen, egituraren ardatz bateko enborraren erdiaren inguruan. Tomek esan zuen Jimen ohearen atzean ginela orduantxe, eta haren azpian egingo genuela zuloa, eta lana amaitzean txabola barruko inor ez zela konturatuko han zuloa zegoenik, Jimen ohe estalkiak ia lurreraino zintzilikatzen baitzuen, eta hura altxa beharko litzateke haren azpian zuloa ikusteko. Zuloa egin eta egin aritu ginen, labainak erabiliz, ia gau erdira arte; gero leher eginik geratu ginen, eskuak babaz beteta genituen, eta ez ziruden oraindik gehiegi aurreratu genuenik. Azkenean esan nuen:
"This ain't no thirty-seven year job, this is a thirty-eight year job, Tom Sawyer."	— Hau ez duk hogeita hamazazpi urtetako lana, hogeita hemezortzi urtetako lana duk, Tom Sawyer.
He never said nothing. But he sighed, and pretty soon he stopped digging, and then for a good little while I knowed he was thinking. Then he says:	Ez zuen ezer esan. Baina hasperen egin zuen, eta tarte batean utzi zion zuloa egiteari, gero tarte luzeagoan pentsatzen ari zela konturatu nintzen. Orduan esan zuen:
"It ain't no use, Huck, it ain't agoing to work. If we was prisoners it would, because then we'd have as many years as we wanted, and no hurry; and we wouldn't get but a few minutes to dig, every day, while they was changing watches, and so our hands wouldn't get blistered, and we could keep it up right along, year in and year out, and do it right, and the way it ought to be done. But we can't fool along, we got to rush; we ain't got no time to spare. If we was to put in another night this way, we'd have to knock off for a week to let our hands get well—couldn't touch a case-knife with them sooner."	— Alferrik duk, Huck, ez diagu aurrera aterako. Presoak gu izan bagina egin ahal izango genuke, nahi adina urte izango genituzkeelako, eta presarik ez; eta ezingo genuke egunean minutu batzuetan besterik zuloa egin, zaintza aldatzen ziaten bitartean, eta horrela eskuak ez litzaizkiguke babaz beteko, eta aurrera egin ahal izango genuke, urte batean eta bestean, eta zuloa ongi egingo genuke, legeak esaten duen bezala. Baina ezin diagu denborarik galdu, azkar ibili beharra diagu; ez diagu denbora alferrik galtzeko astirik. Beste gau batean modu honetan arituz gero, astebetez utzi beharko genioke zulatzeari eskuak sendatu bitartean..., ezingo genuke lehenago labanarik hartu.
"Well, then, what we going to do, Tom? "	— Eta, zer egingo diagu orduan, Tom?
"I'll tell you. It ain't right, and it ain't moral, and I wouldn't like it to get out—but there ain't only just the one way; we got to dig him out with the picks, and let on it's case-knives."	— Esango diat. Ez duk gauza zuzena, ez duk morala, eta ez nuke inork jakitea nahi..., baina bide bat besterik ez zegok; pikotxak erabiliz egin beharko diagu zuloa, eta labainak direla itxura egingo diagu.
"Now you're talking!" I says; "your head gets leveler and leveler all the time, Tom Sawyer," I says. "Picks is the thing, moral or no moral; and as for me, I don't care shucks for the morality of it, nohow. When,! start in to steal a nigger, or a watermelon, or a Sunday-school book, I ain't no ways particular how it's done so it's done. What I want is my nigger; or what I want is my watermelon; or what I want is my Sunday-school book; and if a pick's the handiest thing, that's the thing I'm agoing to dig that nigger or that watermelon or that Sunday-school book out with; and I don't give a dead rat what the authorities thinks about it nuther."	— Orain hitz egin duk behar bezala! —esan nuen—. Gero eta buru argiagoa duk, Tom Sawyer —esan nuen—. Pikotxak dituk gauza zuzena, legezko izan al ez izan; eta niri behintzat, bost axola zaidak horren legezkotasuna. Beltza lapurtzen badut, edo sandia, edo igandeetako eskola liburua, ez zaidak batere axola nola egin, egitea baizik. Beltza nahi diat, edo sandia behar diat, edo igandeetako eskola liburua; eta pikotxa baduk gauzarik eskuragarriena, hori erabiliko diat beltza zulotik ateratzeko, edo sandia, nahiz igandeetako eskola liburua eskuratzeko; eta arratoi hilik ere ez nuke emango autoritateek horretaz zer dioten jakiteagatik.

Figure 7. Paragraph alignment of Chapter 36 between the original English (**left**) and Basque translation (**right**) of the novel "Adventures of Huckleberry Finn".

6. Materials and Methods

Our source code and the relevant corpora used are available online on GitHub at the URL: https://github.com/zzcoolj/rosetta. The online Dashboard is available at the URL: https://rosetta.univ-lille.fr/rosetta-translation-dashboard/.

7. Conclusions

We proposed and experimented a new collaborative and inclusive model for Knowledge Organization Systems (KOS) for sustaining cultural heritage and language diversity. Our first experiment was conducted on translated fictional texts and in particular the world-famous and well-traveled American novel *Adventures of Huckleberry Finn*. This paper reports on 62 metadata and 30 full text collected files including 22 under-resourced languages. The data curation process was based on contributions of end-users as well as scientific and scholarly communities from across borders, languages, nations, continents, and disciplines. Collected translations were processed to build 10 parallel corpora that are sentence-aligned pairs of English with Basque (a European under-resourced language), Bulgarian, Dutch, Finnish, German, Hungarian, Polish, Portuguese, Russian, and Ukrainian, processed out of 30 collected translations. More translations are being collected and processed throughout the project's duration.

Author Contributions: Conceptualization, A.F.; methodology, A.F. and Z.Z.; software, Z.Z. and A.Z.; validation, A.F., R.J. and S.F.F.; formal analysis, A.F., R.J. and S.F.F.; investigation, A.F., S.F.F. and R.J.; resources, S.F.F. and R.J.; data curation, S.F.F., R.J. and A.F.; writing—original draft preparation, A.F.; writing—review and editing, A.F., Z.Z., R.J., S.F.F., P.Z., L.F. and W.M.E.H.; visualization, Z.Z. and A.Z.; supervision, S.F.F., R.J., A.F. and P.Z.; project administration, S.F.F., R.J. and A.F.; and funding acquisition, S.F.F. and R.J.

Funding: This research was funded by the France-Stanford Center For Interdisciplinary Studies in Stanford, USA.

Acknowledgments: This research work was conducted within the framework of the ROSETTA project funded by the France-Stanford Center For Interdisciplinary Studies in Stanford, USA. It is a partnership with Université de Lille(EA 4073-GERiiCO and EA 4074-CECILLE), Stanford University and the Université Paris-Saclay, LIMSI-CNRS.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Otlet, P. *Traité de Documentation: Le livre sur le Livre: Théorie et Pratique*; Mundaneum: Bruxelles, Belgium, 1934.
2. Krauwer, S. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In Proceedings of the International Workshop Speech and Computer, Moscow, Russia, 27–29 October 2003.

3. Arppe, A.; Lachler, J.; Trosterud, T.; Antonsen, L.; Moshagen, S.N. Basic Language Resource Kits for Endangered Languages: A Case Study of Plains Cree. In Proceedings of the 2nd Workshop on Collaboration and Computing for Under-Resourced Languages Workshop (CCURL 2016), Portorož, Slovenia, 23 May 2016; pp. 1–8.
4. Scannell, K. The Crubadan Project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*; University Press Leuven: Louvain-la-Neuve, Belgium, 2007; pp. 5–15.
5. Fraisse, A.; Boitet, C.; Blanchon, H.; Bellynck, V. A Solution for in Context and Collaborative Localization of most Commercial and Free Software. In Proceedings of the 4th Language and Technology Conference (LTC 2009), Poznań, Poland, 6–8 November 2009; pp. 536–540.
6. Fraisse, A.; Boitet, C.; Bellynck, V. An In Context and Collaborative Software Localisation Model. In Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Mumbai, India, 8–15 December 2012; pp. 141–146.
7. Roukos, S.; Graff, D.; Melamed, D. *Hansard French/English*; Linguistic Data Consortium: Philadelphia, PA, USA, 1995.
8. Koehn, P. Europarl: A Parallel Corpus for Statistical Machine Translation. In Proceedings of the Tenth Machine Translation Summit, Phuket, Thailand, 12–16 September 2005; pp. 79–86.
9. Ziemski, M.; Junczys-Dowmunt, M.; Pouliquen, B. The United Nations Parallel Corpus V1.0. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 23–28 May 2016; pp. 3530–3534.
10. Otlet, P. *Monde: Essai d'universalisme: Connaissance du Monde, Sentiment du Monde, Action Organisée et Plan du Monde;* Mundaneum: Brussels, Belgium, 1935.
11. Rayward, W.B. The legacy of Paul Otlet, pioneer of information science. *Aust. Libr. J.* **1992**, *41*, 90–102. [CrossRef]
12. Hudon, M. Multilingual Thesaurus Construction-Integrating the Views of Different Cultures in One Gateway to Knowledge and Concepts. *Inf. Serv. Use* **1997**, *17*, 11–123. [CrossRef]
13. Hudon, M. Accessing Documents and Information in a World without Frontiers. *Index* **1999**, *21*, 156–159.
14. Barát, Á.H. Knowledge Organization in the Cross-Cultural and Multicultural Society. *Adv. Knowl. Org.* **2008**, *11* , 91–97.
15. Resnik, P.; Olsen, M.B.; Mona, D. The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'. *Comput. Humanit.* **1999**, *33*, 129–153. [CrossRef]
16. Mayer, T.; Cysouw, M. Creating a Massively Parallel Bible Corpus. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, 26–31 May 2014; pp. 3158–3163.
17. Christodouloupoulos, C.; Steedman, M. A massively parallel corpus: The Bible in 100 languages. *Lang. Resour. Eval.* **2015**, *49*, 375–395. [CrossRef] [PubMed]
18. Choudhary, N.; Jha, G.N. Creating Multilingual Parallel Corpora in Indian Languages. In *Human Language Technology Challenges for Computer Science and Linguistics*; Springer International Publishing: Cham, Germany, 2014; pp. 527–537.
19. Jha, G.N. The TDIL Program and the Indian Language Corpora Initiative (ILCI). In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta, 17–23 May 2010.
20. Steinberger, R.; Pouliquen, B.; Widiger, A.; Ignat, C.; Erjavec, T.; Tufiş, D.; Varga, D. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th International Conference on Language Resources and Evaluation, Genoa, Italy, 24–26 May 2006; pp. 2142–2147.
21. Cysouw, M.; Walchli, B. Parallel texts: Using translational equivalents in linguistic typology. *Sprachtypol. Univers. STUF* **2007**, *60*, 95–99. [CrossRef]
22. Druskat, S.; Gast, V.; Krause, T.; Zipser, F. corpus-tools. org: An Interoperable Generic Software Tool Set for Multi-layer Linguistic Corpora. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 23–28 May 2016; pp. 4492–4499.

23. Gilmanov, T.; Scrivner, O.; Kübler, S. SWIFT Aligner, A Multifunctional Tool for Parallel Corpora: Visualization, Word Alignment, and (Morpho)-Syntactic Cross-Language Transfer. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland, 26–31 May 2014; pp. 2913–2919.
24. Smith, N.; Jahr, M. Cairo: An Alignment Visualization Tool. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece, 30 May–2 June 2000.
25. Gomes, F.T.; Pardo, T.A.; de Medeiros Caseli, H. Visualtca: Uma ferramenta visual on-line para alinhamento sentencial de textos paralelos. In Proceedings of the Anais do XXVII Congresso da Sociedade Brasileira de Computação-V Workshop em Tecnologia da Informação e da Linguagem Humana (TIL), Rio de Janeiro, 5–6 July 2007; pp. 1729–1732.
26. Fleury, S.; Zimina, M. Exploring Translation Corpora with MkAlign. Available online: https://www.researchgate.net/profile/Maria_Zimina5/publication/49135660_Exploring_Translation_Corpora_with_MkAlign/links/5baa93ab299bf13e604c87eb/Exploring-Translation-Corpora-with-MkAlign.pdf (accessed on 20 September 2019).
27. Teets, M.; Goldner, M. Libraries' Role in Curating and Exposing Big Data. *Future Int.* **2013**, *5*, 429–438. [CrossRef]
28. Cassin, B.; Ducimetière, N. *Les Routes de la traduction. Babel à Genève*; Gallimard: Paris, France, 2017.
29. Fishkin, S.F. DEEP MAPS: A Brief for Digital Palimpsest Mapping Projects (DPMPs) or 'Deep Maps'. Available online: https://escholarship.org/uc/item/92v100t0 (accessed on 20 September 2019).
30. Rodney, R.M. *Mark Twain International: A Bibliography and Interpretation of His Wordwide Popularity*; Greenwood Press: Westport, CT, USA, 1982.
31. Twain, M. *Adventures of Huckleberry Finn*; Charles, L., Ed.; Webster and Company: Hartford, CT, USA, 1885.
32. Fraisse, A.; Jenn, R.; Fishkin, S.F. Parallel Corpora for Under-Resourced Languages Using Translated Fictional Texts. In Proceedings of the LREC 2018 Workshop CCURL2018—Sustaining Knowledge Diversity in the Digital Age, Miyazaki, Japan, 7–12 May 2018; pp. 39–43.
33. Gale, W.A.; Church, K.W. A Program for Aligning Sentences in Bilingual Corpora. *Comput. Linguist.* **1993**, *19*, 75–102.
34. Brown, P.F.; Pietra, V.J.D.; Pietra, S.A.D.; Mercer, R.L. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Comput. Linguist.* **1993**, *19*, 263–311.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).