Building Multilingual Parallel Corpora for Under-Resourced Languages Using Translated Fictional Texts

Amel Fraisse[1], Ronald Jenn[1], Shelley Fisher Fishkin[2]
[1]University of Lille (France), [2]Stanford University (USA)
{amel.fraisse, ronald.jenn}@univ-lille.fr, sfishkin@stanford.edu

Abstract

In this paper, we present an ongoing research 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 fluid and collaborative. It 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 10 parallel corpus that are now chapter aligned pairing English with Arabic, Basque, Bengali, Bulgarian, Dutch, Hungarian, Polish, Russian, Turkish and Ukrainian processed out of a total of 20 collected translations.

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

1. Introduction

Out of the world's 6000+ languages only a small fraction, a dozen or so, currently enjoy the benefits of modern language technologies such as speech recognition or machine translation. A larger but still modest number, 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 and the like (Krauwer, 2003; Arppe et al., 2016). This means that as mentioned by (Scannell, 2007) over 98% of world languages lack most, and usually all, of these language resources. Even for well-endowed languages, parallel corpora are a rare resource. And yet, there is great need for them. Parallel corpora are a valuable resource for linguistic research and natural language processing (NLP) applications. Such corpora are often used for testing new tools and methods in Statistical Machine Translation (SMT), where large amounts of aligned data are often used to learn word alignment models between two languages (Och and Ney, 2003). Building such corpora for endangered languages presupposes the existence of translated language materials in these languages, where there are mostly available in print and awaiting digitization. When translation or software localization does occur it is mostly into commercially important languages (Fraisse et al., 2009; Fraisse et al., 2012; Roukos et al., 1995; Koehn, 2005; Ziemski et al., 2016).

Multilingual online digital libraries and archival projects collect documents and make them available to a wide audience : the Wikisource project [1], an online digital library of free content textual sources, the Internet Archive project[2] building a digital library of Internet sites and other cultural artifacts in digital form such as books and audio records, or the Gutenberg project[3] 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. In this work, we present our research within the framework of the funded Global Huck 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. We conducted a first experiment on the novel "Adventures of Huckleberry Finn" by the American author Mark Twain (1885). This fictional text was chosen because we knew for a fact, thanks to previous scholarship, that it was translated early in many different languages worldwide and that continued interest in the novel throughout the 20th and well into the 21st centuries guaranteed that a great number of translations in sometimes unexpected under-resourced languages were available. What makes the translation of such a fictional text especially valuable for the construction of multilingual parallel corpora is that it uses everyday commonplace words and phrases to describe its actions and plot. It is therefore not confined to a specific domain although the novel does revolve around the universal topics of freedom, slavery, race relations, oppression, emancipation and violence, so many topics that account for its fame and popularity. Our main focus in this paper is the collection and construction of multilingual parallel corpora built thanks to this particular novel with a view to provide digital corpora that will eventually be turned into dictionaries, thesauri, lexicons, and other linguistic resources.

2. Related work

Over the last few years, there has been a growing interest and awareness among the scientific community and lo-

[1]https://wikisource.org
[2]https://archive.org
[3]https://www.gutenberg.org

cally among advocates of minority languages in sustaining and expanding the existing resources in endangered languages and digitizing them in order to preserve and promote knowledge and language diversity.

In particular in relation to parallel corpora for under-resourced languages, some research works focused on religious texts such as the Bible as a relevant source to compile massively parallel corpora (Resnik et al., 1999). This line of research, which entailed 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 (Mayer and Cysouw, 2014). In (Christodouloupoulos and Steedman, 2015), the authors built a massively parallel corpus based on 100 translations of the Bible, emphasizing difficulties in acquiring and processing the raw material.

Kevin Scanell (2007) focused on the creation of web-crawled corpora for many minority and under-resourced languages and the development of open NLP tools for these languages in collaboration with native speakers. In (Choudhary and Jha, 2014; Jha, 2010), the authors created a 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 (Steinberger et al., 2006), 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 nowadays' 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", "Master i Margarita") or translations from the web, mostly available for a set of closely related languages (Cysouw and Walchli, 2007; Mayer and Cysouw, 2014). Most of these texts mainly concern well-endowed largely known languages.

3. 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. As the UNESCO Index Translationum[4] shows the American writer is ranked 15 in the top-50 of the most translated authors worldwide. His works have been translated into almost every language in which books are printed (Rodney, 1982) including under-resourced languages. The novel "Adventures of Huckleberry" (Twain, 1885) is one of the most commonly translated of his books. Rodney (1982) identified 375 translations in 54 different languages as of 1976. As UNESCO's Index Translationum suggests, hundreds of additional translations have been published in the four decades since Rodney completed his survey. Table 1 shows the scores of languages into which the book has been translated. The list includes Afrikaans, Albanian, Arabic, Assamese, Bengali, Bulgarian, Burmese, Catalan, Chinese, Chuvash, Czech, Danish, Dutch, Estonian, Farsi, Finnish, French, German, Georgian, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Kazakh, Korean, Kirghiz, Latvian, Lithuanian, Macedonian, Malay, Malayalam, Marathi, Norwegian, Oriya, Polish, Portuguese, Romanian, Russian, Serbo-Croatian, Sinhalese, Slovak, Slovenian, Spanish, Swedish, Tamil, Tatar, Telugu, Thai, Turkish, Ukranian, and Uzbek. 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.

[4] http://www.unesco.org/xtrans/

4. Copyright and Digitization

Copyright issues are one of the major challenges in digitizing works in print that are still under copyright. According to the Berne convention, the copyright duration is 50 years after the author's death, while local laws extend that duration to up to 70 years. Choosing a text such as "Adventures of Huckleberry Finn", first published in 1885 and so immediately popular that it was translated into many languages means that a range of versions is available in the public domain and therefore readily available for our research.

4.1. Collecting Mark Twain's translation in under-resourced languages

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. Those Twain scholars can be teachers of American studies and/or literature or work in another field but keep an interest in Mark Twain. 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 by Rodney (1982), the Twain community provided us with a compiled list of additional references through, for example, field research at the UNESCO in Paris. The UNESCO has, for many years starting in the late 1920s early 1930s, carried out a yearly survey of translations around the world called the Index Translationum. Additional and even more recent translations of Mark Twain have been discovered within the framework of the Global Huck project. 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. Beside the numerous versions in well-endowed languages such as French, German, Italian and Spanish, the novel was translated into a large number of under-resourced languages (Table 1).

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

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

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. In other cases, such as Bengali for example, the digital version was in image format and could therefore not be processed as such. In this case we transcribed the text following an approach described in the next section of this paper. 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. Within the framework of this project, local institutions are crucial because they have the knowledge, the expertise and they help us determine the copyright status of the versions we deal with. This project therefore enhances language diversity by tapping into the local institutions of under-resourced languages.

4.2. A crowdsourcing approach for text transcription

Over the past few years, many crowdsourced transcription projects have been created in order to transcribe speech, typed or handwritten documents. A wide spectrum of languages, historical periods, and geographic areas are represented by this type of project. For example, the City Archive of Leuven[5] crowdsourced the transcription of more than 950,000 Dutch-language register pages from the Leuven court of Aldermen during the years 1362 to 1795. The Ancient Lives project (Williams et al., 2014) asks online volunteers to transcribe fragment of ancient Greek texts from a Papyri collection. The Rediscovering Indigenous Languages project[6] crowdsourced the transcription of historic word lists, records and other documents relating

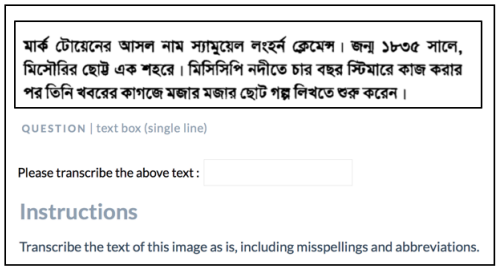

Figure 1: Example of the Bengali transcription task. The digitized image is at the top. Below it is the task instruction.

to indigenous Australian languages. As showed by several research works (Novotney and Callison-Burch, 2010; Gelas et al., 2011; Munyaradzi and Suleman, 2013), crowdsourced indigenous language transcription produces reliable transcriptions of high quality. We used CrowdFlower (Biewald, 2012) an enhanced service that feeds into Amazon's Mechanical Turk[7] and other crowdsourcing systems to transcribe digital versions that came as images, whether from local institutions or collected from the web. It provides convenient management tools that show the performance of workers for a task. The CrowdFlower User Interfaces (UIs) tend to fall into a set of tasks such us selection, categorization, text input or text transcription. The next step is the data import which may be uploaded as CSV or XLS files. Each page of each scanned translation represents a line in these data files which is associated to a task unit on CrowdFlower. The task asked workers to transcribe the text of one page as is, including misspellings and abbreviations. Figure 1 shows the example of the transcription task for the Bengali text.

5. The collected multilingual parallel corpora

In total, we collected digital translations in 20 under-resourced languages : Arabic, Basque, Bengali, Bulgar-

[5]http://itineranova.be/in/home

[6]https://transcripts.sl.nsw.gov.au

[7]http://www.mturk.com

English	Arabic	Basque
You don't know about me, without you have read a book by the name of "Adventures of Huckleberry Finn", 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.	الك لن تعرفنى ايها القارىء الا اذا كنت قد قرات كتابا بعنوان « توم سوير » (!) ، وان كنت اعتقد الا اهمية لذالك ، فقد الف مستر « مارك توين » هذا الكتاب وضمنه جوهر الحقيقة ، ومعلوم سمح لنفسه بان يبتدع بضع حقائق نسجيها خياله ، فانه توخى الصدق بصفة عامة . وعلى اية حال ، فانني لا اذكر انني قابلت انسانا لم يكذب مرة او اخرى ، ولست استثنى من ذالك « الخالة بولى » او « الارملة دوجلاس » وربما « مارى » . ولقد ذكر « مارك توين » كل شيء عن « الخالة بولى » ـ وهى خالة توم ـ ومارى ، والارملة دوجلاس فى هذا الكتاب .. وهو كتاب صادق فى مجموعه مع بعض الجنوح الى الخيال كما قلت من قبل .	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.
Bengali	Dutch	Turkish
মার্ক টোয়েনের আসল নাম স্যামুয়েল লংহর্ন ক্লেমেন্স। জন্ম ১৮৩৫ সালে, মিসৌরির ছোট্ট এক শহরে। মিসিসিপি নদীতে চার বছর স্টিমারে কাজ করার পর তিনি খবরের কাগজে মজার মজার ছোট গল্প লিখতে শুরু করেন। মার্ক টোয়েন অনেক ছোটগল্প ও উপন্যাস লিখেছেন। 'এ কানেকটিকাট ইয়াংকি ইন কিং আর্থারস কোর্ট', 'দ্যা প্রিন্স এন্ড দ্যা পপার', 'লাইফ অন দ্যা মিসিসিপি' ইত্যাদি মার্ক টোয়েনের নামকরা উপন্যাস হলেও যে দুটো বই তাঁকে পৃথিবীখ্যাত লেখকের মর্যাদা দিয়েছে তা হচ্ছে, 'দ্যা এ্যাডভেঞ্চার অব টম সয়ার' এবং 'দ্যা এ্যাডভেঞ্চার অব হাকলবেরি ফিন'। মার্ক টোয়েনের নিজের ভাষায় এ বই দুটিকে ফুটে উঠেছে তাঁর শৈশব জীবনের অনেক কাহিনী। মাত্র বার বছর বয়সে স্কুল ছেড়ে দেয়া মার্ক টোয়েন পরবর্তীতে আমেরিকা ও ইউরোপের বিশ্ববিদ্যালয় থেকে সম্মানসূচক ডিগ্রি পেয়েছেন। মার্ক টোয়েন আমেরিকার লেখক হলেও তাঁর পাঠক শুধু আমেরিকার ভূখন্ডে সীমিত থাকেনি। পৃথিবীর প্রতিটি দেশে রয়েছে তাঁর অসংখ্য গুনমুগ্ধ পাঠক। ১৯১০ সালে এই মহান লেখক মৃত্যুবরণ করেন।	Als gij het boek getiteld 'de lotgevallen van Tom Sawyer' niet gelezen hebt, weet gij niets van mij, maar dat doet er minder toe. Dat boek was geschreven door Mark Twain, en hij vertelde de waarheid, in hoofdzaak althans. Er waren wel dingen, die hij wat opsierde, maar in hoofdzaak vertelde hij de waarheid. Dat hindert niet. Ik heb nog nooit iemand gezien, die niet wel eens loog, of het moest tante Polly, of de weduwe, of des noods ook Mary geweest zijn. Over tante Polly - dat wil zeggen Tom's tante Polly - en over Mary en over de weduwe Douglas, over al die menschen is reeds gesproken in het boek - dat voor het grootste gedeelte een waar verhaal bevat; natuurlijk met eenige uitbreidingen, zooals ik reeds gezegd heb.	Tom Sawyer'ın Maceraları adlı kitabı okumadıysanız beni tanımazsınız, ama ziyanı yok. O kitabı Bay Mark Twain yazdı ve kendisi çoğu yerde doğruyu söylüyor. Kimi şeyleri abartmış, fakat söylediklerı doğru çoğunlukla. Ne çıkar ki zaten. Öyle ya da böyle yalan söylemeyen kimseyi tanımadım, Polly Teyze hariç, bir de dul bayan, belki bir de Mary. Polly Teyze –Tom'un Polly Teyzesi yani– ve Mary, bir de Dul Bayan Douglas'tan o kitapta bahsediliyor; demin söylediğim gibi, kitapta çoğunlukla doğru şeyler söylenmiş ama abartılmış yerler de var.

Table 2: Original text of the first paragraph of Chapter 1 of "Adventures of Huckleberry Finn" Finn and its translations in 5 under-resourced languages.

ian, Chinese (simplified), Chinese (traditional), Croatian, Czech, Dutch, Finnish, Hebrew, Hungarian, Polish, Portuguese, Romanian, Russian, Tamil, Turkish, Ukrainian and Vietnamese. The original version of the novel as well as collected translations are already structured by chapter. To perform corpus alignment, we kept the original structure by putting each chapter between an opening and closing chapter tag (<CHAPTER> </CHAPTER)). Even though we do not have a command of each and every target language, aligning the chapters was facilitated by the fact that all translations had the same number of chapters as the original English version (43 chapters) and each chapter starts on a new page and has a title of its own. In this work, we aligned 10 translations with the English source text : English-Arabic, English-Basque, English-Bengali, English-Bulgarian, English-Dutch, English-Hungarian, English-Polish, English-Russian, English-Turkish and English-Ukrainian. The remaining translations are under processing and will be included in a further version of the corpus. Table 2 shows as an example the original text of the first paragraph of chapter 1 of "Adventures of Huckleberry Finn" aligned with its translations in 5 under-resourced languages: Arabic, Basque and Bengali, Dutch and Turkish.

6. Expected use and availability

The major expected use will be to provide statistical machine translation systems with a rich parallel corpus for under-resourced languages. This current version of our corpus displays 10 under-resourced languages and others are being processed. One of our ultimate goals is to reach out to more less-resourced languages. Another goal is to engage scholars in the field of digital humanities as well as languages and Translation Studies specialists to address a number of fundamental questions. What happens in translation ? What is the impact of the linguistic and cultural transfer of the novel on its textual and iconic nature? An aligned digital corpus would allow them to evaluate the modifications and adaptations set up by translators and the translation process. A stable and reliable corpus will made available to them to conduct their own research. This research work will benefit academics both in their research and teaching activities in various areas of the humanities. More texts will be collected throughout the project's duration and thus the publicly available parallel corpus will be released. In the mean time, a first version of the corpus is available online and accessible on Github at the URL: `https://github.com/GlobalHuck/UnderResourcedLanguagesParallelCorpora/releases/tag/v1.0`

7. Conclusion and future works

In this paper 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. Our first experiment was conducted on the world-famous and well-traveled American novel "Adventures of Huckleberry Finn". This paper reports on 10 parallel corpus that are now chapter aligned pairing English with Arabic, Basque, Bengali, Bulgarian, Dutch, Hungarian, Polish, Russian, Turkish and Ukrainian processed out of a total of 20 collected translations that are being processed. More texts will be collected through the project duration.

Acknowledgments

This research work is conducted within the framework of the Global Huck project funded by the MESHS (Maison Européenne des Sciences de l'Homme et de la Société) in Lille, France. It is a partnership with Stanford University. We are also grateful to the Center for Mark Twain Studies in Elmira, N.Y.

8. Bibliographical references

Arppe, A., Lachler, J., Trosterud, T., Antonsen, L., and Moshagen, S. N. (2016). Basic language resource kits for endangered languages: A case study of plains cree. In *Proceedings of the the 2nd Workshop on Collaboration and Computing for Under-Resourced Languages Workshop (CCURL 2016)*, pages 1–8, Portorož, Slovenia, may 23.

Biewald, L. (2012). Massive multiplayer human computation for fun, money, and survival. *Current Trends in Web Engineering*, pages 171—-176.

Choudhary, N. and Jha, G. N. (2014). Creating multilingual parallel corpora in indian languages. In *Human Language Technology Challenges for Computer Science and Linguistics*, pages 527–537, Cham. Springer International Publishing.

Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Cysouw, M. and Walchli, B. (2007). Parallel texts: Using translational equivalents in linguistic typology. *Sprachtypologie und Universalienforschung STUF*, 60(2):95–99.

Fraisse, A., Boitet, C., Blanchon, H., and Bellynck, V. (2009). 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)*, pages 536–540, Poznań, Poland., november 6-8.

Fraisse, A., Boitet, C., and Bellynck, V. (2012). An in context and collaborative software localisation model. In *proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, page 141–146, India, Mumbai., December 16-18.

Gelas, H., Abate, S., Besacier, L., and Pellegrino, F. (2011). Quality assessment of crowdsourcing transcriptions for african languages. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, pages 3065–3068, Florence, Italy, august 27-31.

Jha, G. N. (2010). The tdil program and the indian language corpora initiative (ilci). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 19-21.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit 2005*, september 12-16.

Krauwer, S. (2003). 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, october.

Mayer, T. and Cysouw, M. (2014). Creating a massively parallel bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, may 26-31.

Munyaradzi, N. and Suleman, H. (2013). Quality assessment in crowdsourced indigenous language transcription. *Research and Advanced Technology for Digital Libraries.TPDL 2013. Lecture Notes in Computer Science*, 8092:13–22.

Novotney, S. and Callison-Burch, C. (2010). Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 207–215, Los Angeles, California, june 2-4.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Resnik, P., Olsen, M. B., and Mona, D. (1999). The bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1):129–153.

Rodney, R. M. (1982). *Mark Twain International: A Bibliography and Interpretation of his Wordwide Popularity*. Greenwood Press, Westport, CT.

Roukos, S., Graff, D., and Melamed, D. (1995). Hansard french/english. In *Philadelphia: Linguistic Data Consortium*.

Scannell, K. (2007). The crubadan project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, pages 5–15, Louvain-la-Neuve, Belgium.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., and Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 2142–2147, Genoa, Italy, may 24-26.

Twain, M. (1885). *Adventures of Huckleberry Finn*. Charles L. Webster and Company, Hartford, Connecticut.

Williams, A. C., Wallin, J. F., Yu, H., Perale, M., Carroll, H. D., Lamblin, A.-F., Fortson, L., Obbink, D., Lintott, C. J., and Brusuelas, J. H. (2014). A computational pipeline for crowdsourced transcriptions of ancient greek papyrus fragments. In *Proceedings of the International Conference on Big Data, IEEE Big Data 2014*, pages 100–105, Washington, United States, october 27-30.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, page 3530–3534, Portorož, Slovenia, may 24-26.