# Crowdsourcing Model for Multilingual Corpus and Knowledge Construction: The Case of Transnational Mark Twain

Amel Fraisse, Ronald Jenn, Quoc-Tan Tran

# Crowdsourcing Model for Multilingual Corpus and Knowledge Construction: The Case of Transnational Mark Twain

Amel Fraisse
*GERiiCO, University of Lille, France*
Ronald Jenn
*CECILLE, University of Lille, France*
Quoc-Tan Tran
*GERiiCO, University of Lille, France*

**Abstract**

**Purpose/Thesis:** We describe a new approach that addresses key challenges to multilingual corpus by merging collective human intelligence (crowdsourcing) and automated knowledge construction and extraction methods in a symbiotic fashion.
**Approach/Methods:** We use a crowdsourcing model to collect and annotate translations of the same literary text.
**Results and conclusions:** The model promotes a dynamic approach to archives that increases the impact of traditional research by presenting the text from a new angle, accessible to a global public.
**Practical implications:** The Global Huck project proposes a new paradigm to assess the contribution of crowdsourcing-based models for collection and annotation purposes.
**Originality/Value:** Choosing the translations of a novel as a field of study is a truly transnational and multilingual collaborative endeavor allowing us to increase our capacity to collect and organize data on a broad, transnational and multilingual scale.

**Keywords**

Deep mapping. Humanities crowdsourcing. Multilingual corpus. Parallel text processing. Under--resourced languages.

*Received: 9 May 2018. Reviewed: 7 June 2018. Revised: 22 June 2018. Accepted: 10 July 2018.*

## 1. Introduction

At a time when databases and digitized archives are growing exponentially, their usefulness to researchers and the public remains limited. Databases and online archives tend to be populated by data that was previously digitally indexed. The rapid growth of crowdsourcing challenges established approaches to knowledge construction and extraction. In fact, whether human expert or machine-based models, they face significant issues either at the problem-solving level (workflow, control, design) or the corpora level (large quantity, noise, multilinguality, natural language problems) (Sabou et al., 2013; Wohlgenannt et al., 2016; Zhang et al., 2011).

In this paper we describe a new approach that addresses the shortcomings by merging collective human intelligence (crowdsourcing) and automated knowledge construction and extraction methods in a symbiotic way. Our approach takes place in the wake of the success of models based on mass participation known as crowdsourcing (Fraisse & Paroubek, 2015; Howe, 2008; Law et al., 2017; Sabou et al., 2012). The combined approach will allow us to take up the challenge presented by voluminous and multilingual corpora, thus contributing to a broadening and an enrichment of our current vision of knowledge-building and knowledge-extraction.

This crowdsourcing model is the core of our project Global Huck, whose aim is to create an innovative, enterprising and stimulating intervention in literary studies, translation studies, and digital humanities. Our mission is exploring domains and territories recently still uncharted and still awaiting collection, interpretation, indexing, mapping, and mining. The type of corpus we focus on is literary translation. There is no existing structure allowing one to survey and assess all the existing translations of the same text in a multitude of languages worldwide. Approaches based on human expertise alone and those based exclusively on machines can both be limited when it comes to working on corpora in a broad range of languages (Christodouloupoulos & Steedman, 2015). Our starting point is to collect and annotate all the existing translations (including in poorly-endowed languages) of a single novel – Mark Twain's *Adventures of Huckleberry Finn* – before visualizing those different versions on an online platform with an interactive map as a point of access. The ultimate goal of this platform is to bring together colleagues from a range of disciplines and create a transnational dialogue around translations of literary works.

## 2. Digital humanities and crowdsourcing

### *2.1. Crowdsourcing as a method of data collection*

As the combination of "crowd" and "outsourcing," "crowdsourcing" describes online projects that involve free or inexpensive labor provided by Internet users around the world. The term refers to the distribution of work when the crowd responds to a call in which contributions are sought to achieve a specific goal. In recent years, this approach has been used in various digital humanities domains such as digital pedagogy, digital curation, and digital scholarship, among others. Generically it could be applied to any discipline, from geography, archaeology, public history to disaster management (Porto de Albuquerque et al., 2016), as well as being associated with a wide variety of scenarios (Prpić et al., 2015). Its emergence has been possible thanks to the evolution of social media technologies and scholarly information infrastructure that have enabled communities, both scientific and public, to collaborate and exchange information.

Transcribe Bentham is a highly recognized and rewarding project both in substance and in form. Being one of the pioneer initiatives in crowdsourced transcription, this project is an invitation to reflect on new approaches to research. One can particularly observe the collaboration between the specialist and the non-specialist (a partial blending of "sacred-profane" cultures) towards common aims: foster innovation, advance knowledge and serve the public. In the field of digital classics, the Ancient Lives project proposes

the collaborative transcription of ancient Greek fragments from the Graeco-Roman era to identify and make them available to researchers. Being part of the Zooniverse network, Ancient Lives is a stimulating case of "citizen cyberscience" whose aims are to involve the volunteers in science and establish a distributed community of citizen scientists (Shuttleworth, 2016).

To date, it appears that digitization initiatives have embraced the crowdsourcing techniques to draw in the skills of communities or individuals to transcribe, correct and index a predetermined collection of materials. However, as this approach is based primarily on mechanized micro-tasks, "learning practices remain unchanged: correction and transcription, contextualization, classification, co-curation"; therefore, "neither the nature of tasks nor their division in a mechanized process, constitutes the element that changed humanities production of knowledge in the virtual world" (Favier, 2016, 14).

On the other hand, a new feature of crowdsourcing has emerged recently and could cause a changing paradigm: content creation. A prime example is the Great War Archive, funded under the UK's JISC Digitisation Programme. Launched in 2008, it invited the general public to contribute items in private ownership, including letters and diaries, recordings of interviews, and physical artifacts from the Great War period. Public responses were higher than expected and the project had to create a Flickr group to accommodate a great deal of material coming in from the public (Marchionni, 2009). Not only did "this increase users' interest and engagement in the digital resource in the UK, but also brought the initiative to a wider audience" (Coutts, 2016, 89). The content became part of the "Europeana 1914–1918" collection whose idea is to bring together memorabilia and stories across the Continent.

### *2.2. Advocacy and criticism*

Digital humanities (DH) represent a set of values converged among the humanities and the network cultures. Even though at the core these two cultures share a common aim, i.e., to foster innovation, advance knowledge, and serve the public, there are some conflicts of values that generate ongoing debates (Gold, 2012). For the humanities, it is important to defend the knowledge of experts, academic career, and professionalization (Levine et al., 1989). Such a focus on specialization and professional authority will come into conflict with the collaborative, crowdsourcing approaches in the DH. As remarked by Ayers (2004), the academic culture and the IT culture often clash, because the nature of computing has a transparent, dynamic, unstable character, and all the works are carried out by anonymous teams, while the academic culture is defined by critical thinking, debate, balancing innovation and tradition, which is stable and centered on professional identity. As a bridge between these two communities, the DH promote the new set of values by pursuing "a public role for scholarship" (Gold, 2012, 20) through the creation of freely accessible digital archives or by supporting the discussion and the democratic sharing of ideas.

On the other side, there are some criticisms on collaborative and crowdsourcing approaches. Often raised is the question regarding the intellectual authority and quality of the content: how reliable is the evaluation by the crowd in relation to the expert assessment? Advocating "the wisdom of crowds", Surowiecki (2005) discusses how much better decisions a group can make. The author postulates that the average evaluations of a non-specialist

group can be more precise than the evaluation of a handful of experts as individual biases might cancel each other out. However, DH practitioners are more likely to be skeptical. In the Bentham Project, whose primary purpose is to create a scholarly digital edition of Jeremy Bentham's complete works, transcription is an activity that requires a university background to understand the nuances or the relevance of the text. A validation process was then put in place by the project managers. The text is transmitted to the editing service after having been studied by a large number of users; thus, the comparison of the different transcriptions makes it possible to obtain a reliable result. Moreover, not all manuscripts of Bentham are present on the site; the manuscripts with a more difficult level of writing are transcribed directly by the researchers.

There are also ongoing debates around societal issues related to the crowd labor activities (Flinn, 2010). Over the past few years, crowdsourcing has been used to perform a wide variety of challenging tasks for computers, but solvable, such as image recognition, entity resolution, and sentiment analysis, as in the case of Amazon's Mechanical Turk. The principle of mechanized labor is that volunteers choose to realize small tasks called Human Intelligence Task (HIT). In return, they are paid a small amount of money. Some are concerning the ethical nature of cases where the practice of crowdsourcing can deprive other people of a paid job. Then there are questions about legal validity, the abuse of workers' rights, and further implications for the future of work and technology. Cardon and Casilli's remarks on "digital labor" help us see the broad picture:

> [the idea of digital labor] goes through a denunciation of the growing precariousness of content producers, faced with the commodification of their contributions. What type of wage pressure is exercised in the most diverse sectors (journalism, cultural industries, transport, etc.) by the creation of a reserve army of 'ignorant workers', convinced that they are more like consumers, or even beneficiaries of free online services? (Cardon & Casilli, 2015, 16, our own translation).

## 3. Global Huck – A multilingual parallel corpus

### *3.1. Transnational Mark Twain*

Global Huck is a collaborative and transnational digital project between the University of Lille and Stanford University, allowing digital archives and scientific literature worldwide to converse. In these two universities, there are already projects dealing with the issues of DH such as eBalzac (generic and hypertext study on the works of Balzac), Les Monuments aux Morts (exhaustive inventory of war memorials for France and Belgium) in Lille, Mapping the Republic of Letters (repository for metadata on early-modern scholarship), or Spatial History Project (creative spatial, textual and visual analysis) in Stanford. Global Huck is a continuation the already undertaken research and brings together the knowledge and expertise in DH, Translation Studies, and American Transnational Studies.

The "transnational turn" in American Studies has come to dominate the field since Shelley Fisher Fishkin developed the concept in 2004. Transnational work, however, is most often pursued by individual scholars writing in English. Our project intends to raise the bar, demonstrating the potential impact of a blend of crowdsourcing and collaboration across multiple languages and cultures (Fig. 1). Its goal is developing a methodology

for tracking the global circulation of any literary text. By utilizing crowdsourcing and the efforts of collaborating scholars who are committed to the project, Global Huck will collect and annotate translations of Mark Twain's *Adventures of Huckleberry Finn* and scholarship about those translations. Global Huck attempts to break new ground in two ways: (1) It aims to increase our capacity to collect and organize data in the humanities on a broad, transnational, and multilingual scale; and (2) it explores a new paradigm to assess the contribution of crowdsourcing-based models for collection and annotation purposes.

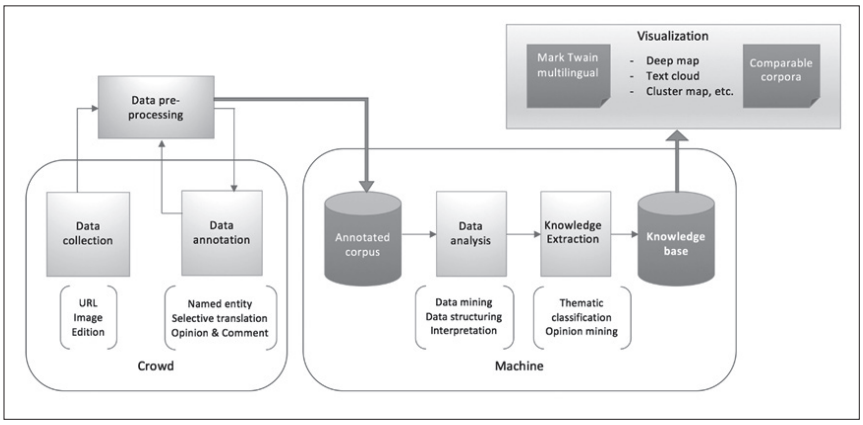


Fig. 1. High-level model for merging crowdsourcing and computational approaches

In fact, the writings of American author Samuel Langhorne Clemens, a.k.a Mark Twain (1835–1910) in English have already been very widely digitized, and large portions of his writings are now in the public domain, available via the Gutenberg project, Internet Archive and elsewhere. Because of the language barrier, the translations of his works remain largely unexplored except in occasional academic articles by Twain scholars. *Adventures of Huckleberry Finn* is the pilot text to be used in the project because of the many methodological advantages and intriguing challenges it presents. Published in 1885, the original text, as well as many of its translations, are now in the public domain what facilitates text manipulation and data mining. Previous scholarship has shown that for ideological as well as literary reasons *Adventures of Huckleberry Finn* has been widely translated around the world (Harrington & Jenn, 2017; Ishihara, 2005; Lai-Henderson, 2015; Rodney, 1982). Fishkin (2010; 2015) claims the ubiquity of Mark Twain's global presence and widespread acclaim. His novel *Adventures of Huckleberry Finn* deals with transnational and universal topics such as slavery, freedom, childhood, racism, and coming of age; this focus, combined with the astounding number of translations available, make it an ideal text to use for the prototype in an investigation of the global circulation of a literary text.

### *3.2. Compiling the multilingual corpus using crowdsourcing*

The collection of the corpus of different existing translations takes place in two steps. Our very first step is to collect the English source text of *Adventures of Huckleberry Finn* and its translations into some well-endowed languages (such as French, German, or Spanish).

Those texts make up the initial core of our study. They have been collected using the open databases offered by the respective national libraries or national archives. We also delve into other digital libraries of free content textual sources to collect the greatest number of translations (whether the whole text or selected passages) that are already in the public domain and in the languages in which scholars involved in the project have expertise. In the second step, we call on the crowd to enrich and build up the rest of the corpus with translations that could not be accessed otherwise. The idea is to call on contributors around the world, via crowdsourcing platforms, to collect and annotate translations of the novel that are available in their mother tongues. In this initial stage of data entry, paid contributors via CrowdFlower enter into tabular form data including languages of translations, translators, publishing houses, URLs for these translations, URLs for the covers, bibliographic records, articles related to the translations (Fig. 2).

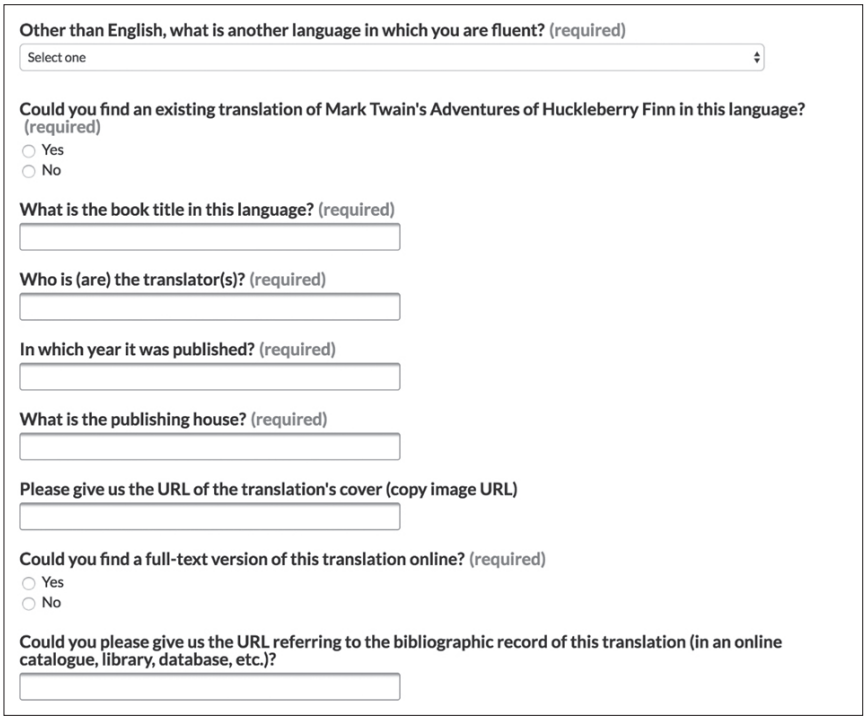


Fig. 2. Questions for each task

The second stage, concerning the annotation and the collation of annotation data for the various translations collected, will involve mainly the expert scholars connected with the project using the CrowdCrafting platform. Precision is vital in both of these stages, and the combination of CrowdFlower to gather the initial materials and CrowdCrafting to expertly curate them will help avoid typos that would crop up in working with optical character recognition, or through other less structured means. Ultimately, a broad range of annotations will be gathered through these two crowdsourcing platforms. Some may deal with illustrations and cover images. Other annotations may provide information on

the names and biographies of the translators involved, and the profile of the publishing houses (area of specialization, status, etc.). Annotations will also note distinctive elements of the translations of the three or four key passages of the novel that will be culled for every translation when possible.

Using the crowd will allow us to obtain and to identify translations in a broad range of languages. Utilizing experts around the world with an in-depth knowledge of literature in these languages will allow them to share their understanding of why translators made their choices, of the cultural work that these translations do in the countries in which they were produced, and of the particular translation challenges the book poses in specific languages. Dealing with a specific novel, this project has defined boundaries in space and time, yet the proposed approach is generic and could be applied and transferable to other types of documents. Many translations, which have not been digitized yet, will have to be collected, indexed and linked together.

### *3.3. Visualizing the data: Concept of "Deep Mapping"*

Our vital aim is to improve data exploration and visualization. The user has a point of access to different translations of the chosen novel on an interactive map that allows for easy navigation between the different versions (see the interface mockup in Fig. 2). Indeed, the interface enables the user to visualize the existing translations by language and country. Particular passages could be compared across multiple translations. The user chooses a point on the world map to visualize the title of the novel in the selected language. S/he can then have access to the entire text in one click when it is in the public domain. The interface also allows the visualization of predefined translated passages as samples. In the specific case of *Adventures of Huckleberry Finn*, it will be "The Notice," the opening and closing paragraphs of the novel as well as other selected excerpts. For copyright reasons, in case when the version in question is not in the public domain yet, the aggregate length of these excerpts will not exceed the fair use limit. The same extract will be available in several languages in the form of parallel corpora. The cover and some illustrations of each translation will be made available, too.

It will also be possible to have access to critical literature about these translations in the respective languages (or in English when available) with a view to evaluate the influence or reputation of one particular translation. Annotations in English help make the choices that translators made in various languages more accessible to users since they will link to (and excerpt from) existing critical commentaries on these translations. Although the materials will be crowdsourced from a broad global public, experts in particular languages will curate the materials to appear on the website.

The ultimate dimension of the project is to put all these documents and archives in conversation by the establishment of a series of links between them. Those links can be semantic and iconographic, making up a fluid and user-friendly whole. Data visualization and access being a major challenge, Global Huck will set up an online platform that will provide a mapping of the different types of resources that will be produced assuming the form of an interactive map (Fig. 3) based on the Deep Maps model put forward by Shelley Fisher Fishkin (2011).

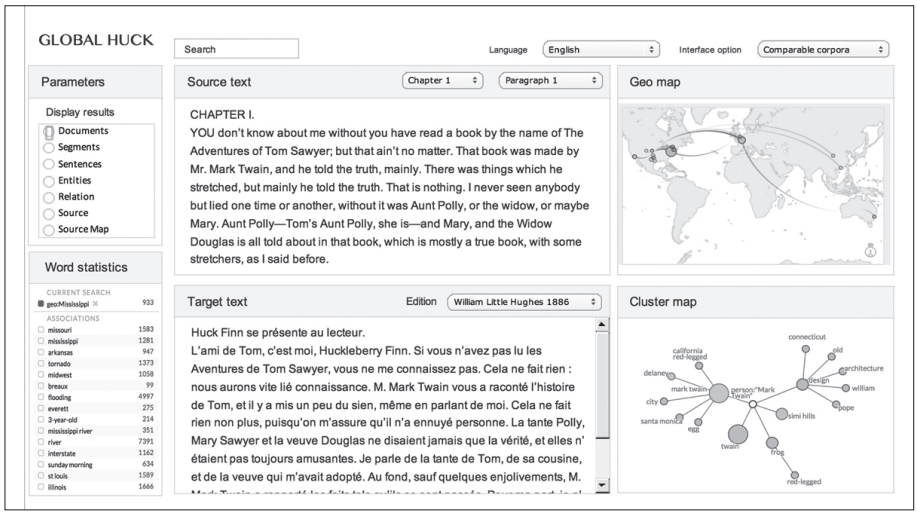


Fig. 3. Global Huck's interface (mock-up)

By Deep Map we mean the development of a platform whose design and approach will be oriented towards the user and whose interface is a map. It is designed to be both interactive and user-friendly:

– Interactive: the user can interactively explore and navigate the platform to access the different types of resources produced. By clicking on a geographical point on the map, the user has access to the translated version(s) of *Huckleberry Finn* found in that specific place/country as well as the metadata of those translations (translator, year, publishing house, comments, illustrations).
– User-friendly: The Global Huck project platform promotes a dynamic approach to archives that allows people to interact in different languages. It increases the impact of traditional research by presenting it in a context and from a new angle, accessible to a global public. The fluidity and interactivity of this intensely collaborative project aim to attract the attention of students and the general public by bringing research alive. It superimposes different approaches to a text in the manner of a palimpsest. Given the fact that many of these translations across different languages are already in invisible conversation with each other, Global Huck is expected to facilitate a new kind of research: for example, Spanish translations of the novel were sometimes made from French translations rather than the original English; Chinese and Korean translations were sometimes made from Japanese translations rather than the original English.

## 4. Conclusion

In this article, we first discussed the context of crowdsourcing as a method of data collection and outlined how crowdsourcing is used to advance research in digital humanities. Then we described the experimental model for multilingual corpora by which we conduct

the Global Huck project. The first aim is to improve the capacity for data collection and organization on a broad, transnational and multilingual scale that involves many languages and cultures including under-resourced languages. Indeed, for these languages crowdsourcing will considerably reduce the costs generated by the scarcity of linguistic and human resources. The second aim of the project is to set up a methodology for tracking the global circulation of any literary text. To facilitate access to the data produced and the visualization of knowledge, Global Huck aims to create and nurture an online generic platform using an interactive map (Deep Map) as an access point. The project is original in that it analyzes and processes data related to the many translations of the same novel. Choosing the translations of a novel as a field of study is a truly transnational and multilingual collaborative endeavor allowing us to increase our capacity to collect and organize data on a broad, transnational, and multilingual scale. A dynamic and large-scale annotated corpus will provide a new way to assess the contribution of crowdsourcing-based models for collection and annotation purposes.

## Acknowledgments

This paper results from an ongoing research developed with the support of the European Center for the Humanities and Social Sciences in Lille (MESHS) as part of the Global Huck project.

## References

Ayers, E. L. (2004). The Academic Culture & The IT Culture: Their Effect on Teaching and Scholarship. *EDUCAUSE Review*, 39(6), 48–62.

Cardon, D., Casilli, A. (2015). *Qu'est-ce que le digital labor?* Bry-sur-Marne: Ina éditions.

Christodouloupoulos, C., Steedman, M. (2015). A Massively Parallel Corpus: The Bible in 100 Languages. *Language Resources and Evaluation*, 49(2), 375–395.

Coutts, M. (2016). *Stepping Away from the Silos: Strategic Collaboration in Digitisation*. Cambridge, MA: Chandos Publishing.

Fraisse, A., Paroubek, P. (2015). Vers des pratiques collaboratives pour les systèmes d'organisation de connaissances. In : E. Chevry-Pébayle (ed.). *Actes du 10ème colloque international du Chapitre français de l'ISKO – Systèmes d'organisation des connaissances et humanités numériques* (289–301). London: ISTE Editions, 289–301.

Favier, L. (2016). Humanities Crowdsourcing. *Zagadnienia Informacji Naukowej. Studia informacyjne*, 54(2), 7–21.

Fishkin, S. F. (2010). *The Mark Twain Anthology: Great Writers on His Life and Works*. New York: Library of America.

Fishkin, S. F. (2011). "Deep Maps": A Brief for Digital Palimpsest Mapping Projects (DPMPs, or "Deep Maps"). *Journal of Transnational American Studies*, 3(2), 1–31.

Fishkin, S. F. (2015). Transnational Mark Twain. In: S. Yuan & D. E. Pease (eds.), *American Studies as Transnational Practice: Turning toward the Transpacific* (109–137). Hanover, New Hampshire: Dartmouth College Press.

Flinn, A. (2010). An Attack on Professionalism and Scholarship? Democratising Archives and the Production of Knowledge. *Ariadne* [online], 62, [29.06.2018], http://www.ariadne.ac.uk/issue62/flinn/

Gold, M. K. (2012). *Debates in the Digital Humanities*. Minneapolis, MN: University of Minnesota Press.

Harrington, P., Jenn, R. (2017). *Mark Twain & France: The Making of a New American Identity*. University of Missouri Press.

Howe, J. (2008). *Crowdsourcing: How the Power of the Crowd is Driving the Future of Business*. London: Random House Business.

Ishihara, T. (2005). *Mark Twain in Japan: The Cultural Reception of an American Icon*. Columbia: University of Missouri Press.

Lai-Henderson, S. (2015). *Mark Twain in China*. Stanford: California Stanford University Press.

Law, E., Gajos, K. Z., Wiggins, A., Gray, M. L., Williams, A. (2017). Crowdsourcing as a Tool for Research: Implications of Uncertainty. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW '17*, 1544–1561.

Levine, G. L., Brooks, P., Culler, J., Garber, M., Kaplan, E. A., Stimpson, C. R. (1989). *Speaking for the Humanities*[online]. American Council of Learned Societies, Occasional Paper No. 7, [29.06.2018], http://archives.acls.org/op/7_Speaking_for_Humanities.htm

Marchionni, P. (2009). Why Are Users So Useful? User Engagement and the Experience of the JISC Digitisation Programme. *Ariadne* [online], 61, [29.06.2018], http://www.ariadne.ac.uk/issue61/marchionni

Porto de Albuquerque, J., Herfort, B., Eckle, M. (2016). The Tasks of the Crowd: A Typology of Tasks in Geographic Information Crowdsourcing and a Case Study in Humanitarian Mapping. *Remote Sensing* [online], 8(10), 859, [29.06.2018], http://www.mdpi.com/2072–4292/8/10/859

Prpić, J., Shukla, P. P., Kietzmann, J. H., McCarthy, I. P. (2015). How to Work a Crowd: Developing Crowd Capital through Crowdsourcing. *Business Horizons*, 58(1), 77–85.

Rodney, R. M. (1982). *Mark Twain International: A Bibliography and Interpretation of His Worldwide Popularity*. Westport, Conn.: Greenwood Press.

Sabou, M., Bontcheva, K., Scharl, A. (2012). Crowdsourcing Research Opportunities: Lessons from Natural Language Processing. In: *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies, i-KNOW 2012 (October 19,2012)*. ACM International Conference Proceeding Series.

Sabou, M., Scharl, A., Michael, F. (2013). Crowdsourced Knowledge Acquisition: Towards Hybrid--Genre Workflows. *International Journal on Semantic Web and Information Systems*, 9(3), 14–41.

Shuttleworth, S. A. (2016). Old Weather: Citizen Scientists in the 19th and 21st Centuries. *Science Museum Group Journal* [online], 3(3), [29.06.2018], http://journal.sciencemuseum.ac.uk/browse/issue-03/old-weather/

Surowiecki, J. (2005). *The Wisdom of Crowds*. New York: Anchor Books.

Wohlgenannt, G., Sabou, M., Hanika, F. (2016). Crowd-Based Ontology Engineering with the Ucomp Protege Plugin. *Semantic Web*, 7(4), 379–398.

Zhang, H., Horvitz, E., Miller, R. C. R., Parkes, D. D. C. (2011). Crowdsourcing General Computation. In: *Proceedings of the 2011 ACM Conference on Human Factors in Computing Systems: May 7–12, 2011, Vancouver, British Columbia*. New York, NY: Association for Computing Machinery.

# Crowdsourcingowy model wielojęzycznego korpusu tekstów literackich i bazy wiedzy: studium przypadku transnarodowej twórczości Marka Twaina – projekt The Global Huck

**Abstrakt**

**Cel/Teza:** Przedstawiono nowe podejście do tworzenia wielojęzycznych korpusów tekstów literackich, które w sposób symbiotyczny łączy zbiorową inteligencję ludzką (crowdsourcing) i zautomatyzowane metody tworzenia baz wiedzy i ekstrakcji informacji oraz kluczowe problemy wynikające z takiego ujęcia.
**Koncepcja/Metody badań:** Zastosowano model crowdsourcingu do zbierania i komentowania różnych przekładów tego samego tekstu literackiego i adnotacji w nich występujących.
**Wyniki i wnioski:** Przedstawiony model reprezentuje dynamiczne podejście do archiwów cyfrowych i pozwala na udoskonalenie tradycyjnych badań (literaturoznawczych) poprzez możliwość prezentacji tekstu literackiego z nowej perspektywy – jako dzieła dostępnego globalnie.
**Zastosowanie praktyczne:** Model systemu dla projektu The Global Huck oferuje nowy paradygmat badań nad cyfrową kolekcją tekstów literackich wraz z ich adnotacjami w ujęciu crowdsourcingowym.
**Oryginalność/Wartość poznawcza:** Badania nad przekładami dzieł literackich mają charakter transnarodowych i kolektywnych procesów badawczych, które pozwalają poszerzyć nasze możliwości pozyskiwania i organizacji tego typu informacji na skalę globalną.

**Słowa kluczowe**

Crowdsourcing w humanistyce. Deep mapping. Języki o ograniczonych zasobach. Korpusy wielojęzyczne. Przetwarzanie równoległe tekstów.

*AMEL FRAISSE, PhD, is Associate Professor at the University of Lille working on Digital Humanities and Natural Language Processing. Her research interests include information extraction and knowledge representation. Her research addresses both methods and applications of text analysis, ranging from collection and exploration of representation models and their cross-language adaptability to the integration of representation frameworks to extract and visualize new knowledge from text.*

*Contact to the Author:*
*amel.fraisse@univ-lille3.fr*
*GERiiCO*
*Domaine Universitaire du Pont de Bois*
*59650 Villeneuve-d'Ascq, France*

*RONALD JENN is Full Professor of Translation and Translation Studies at the University of Lille. He is also a Mark Twain specialist with a focus on the translation and reception of the author in French and in France.*

*Contact to the Author:*
*ronald.jenn@univ-lille3.fr*
*CECILLE*
*Domaine Universitaire du Pont de Bois*
*59650 Villeneuve-d'Ascq, France*

*QUOC-TAN TRAN holds a Master of Research in Library and Information Science from the University of Lille in 2016. His areas of investigation include ethics and cultural interoperability in knowledge organization, participatory methods and Internet ecologies of open knowledge.*

*Contact to the Author:*
*quoc-tan.tran@etu.univ-lille3.fr*
*GERiiCO*
*Domaine Universitaire du Pont de Bois*
*59650 Villeneuve-d'Ascq, France*