

Multi-Modal Citizen Science: From Disambiguation to Transcription of Classical Literature

MARYAM FORADI*, Leipzig University

JAN KASSEL*, Leipzig University

JOHANNES PEIN*, Leipzig University

GREGORY R. CRANE*, Tufts University

The engagement of citizens in the research projects, including Digital Humanities projects, has risen in prominence in the recent years. This type of engagement not only leads to incidental learning of the participant, but also indicates the added value of corpus enrichment via different types of annotations undertaken by users generating the so-called smart texts. Our work focuses on the continuous task of adding new layers of annotation to Classical Literature from the around the world. We aim to provide more extensive tools for readers of smart texts, enhancing their reading comprehension and at the same time empowering the language learning by introducing intellectual tasks, i.e. linking, tagging, and disambiguation. The current study adds a new mode of annotation, audio annotations, to the extensively annotated corpus of poetry by the Persian poet Hafiz, proposing tasks with three different difficulty levels to estimate users' ability in order to rate their annotation in further stages of the project, where no ground truth is available. Annotators with no knowledge of Persian are able to add annotations to the Persian source.

CCS Concepts: • **Information systems** → **Crowdsourcing**; • **Applied computing** → **Interactive learning environments**; • **Human-centered computing** → *User studies*.

Additional Key Words and Phrases: smart texts; citizen science; user experience design; computer assisted language learning (CALL)

1 INTRODUCTION

The increasingly growing adoption of Citizen Science (CS) in academic projects by not only the researchers from STEM disciplines but also (digital) humanists acceleratingly attracts non-professionals from all ages and skills. In the humanities this contribution, with tasks such as text transcription, corrections, classifications, tagging and other annotations, leads to an enrichment of digitized or digitally-born corpora with different types of annotations. However, in most of the CS projects education is not considered a high-priority goal, but rather a side-effect that happens out of the context of the project [17, 26]. On the opposite hand, projects labeled as education projects aim at achieving educational results while the scientific output is considered as a secondary goal [28].

Although crowdsourcing in Natural Language Processing (NLP) profits from the engagement of the members of public as cheap labourers finishing micro-tasks, the cognitive engagement of them is mostly underestimated. The motivation for participation in such projects is often limited to the micro-payments

*All four authors contributed equally to this research.

Authors' addresses: Maryam Foradi, maryam.foradi@uni-leipzig.de, Leipzig University, Augustusplatz 10, Leipzig, Germany, 04109; Jan Kassel, jan.kassel@studserv.uni-leipzig.de, Leipzig University, Augustusplatz 10, Leipzig, Germany, 04109; Johannes Pein, johannes.pein@studserv.uni-leipzig.de, Leipzig University, Augustusplatz 10, Leipzig, Germany, 04109; Gregory R. Crane, gregory.crane@tufts.edu, Tufts University, Medford, USA, MA 021554109.

offered via scholars on platforms such as Amazon Mechanical Turk¹, which can lead to a demographic bias effecting the accuracy of the data [22]. Yet, the recently introduced concept of learnersourcing [15] is an endeavour to provide an opportunity for annotating the learning materials by learners for the other learners and increasing the learning potential of learnersourcers.

In this paper we propose a novel research idea alongside a prototype application² to investigate the extent to which learnersourcing facilitates the phonetic transcription of Persian corpora so that not only the learnersourcers can practice their listening skills, but also other learners have access to more accurate phonetic transcriptions by the means of this medium.

One of the most challenging issues in the digital processing of Arabic and Persian corpora are short vowels (/æ/, /e/ and /o/) that are spoken but not explicitly written. This results in various words that have the same written appearance, but possess entirely different pronunciations, meanings or grammatical features [23]. In this context, the problem is not limited to the automatic phonetic transcription, in which the absence of short vowels in the replacement of Arabic characters with Latin characters leads to the poor readability of transliterated text. Even the existence of a data-set including words written in original language linked with their manual phonetic transcription is deceptive, because the correct form of transcription cannot be identified unless within the context.

While methods of machine transliteration between Arabic and English exist [13], we propose a different, learnersourced approach to this problem. By means of already existing audio recordings of classical texts linked to the source text, we aim at a) choosing the correct form of phonetic transcription among the existing manually transliterated options, and b) adding the short vowels to the automatic transcriptions, where the short vowels are missing. Although the recordings shall be produced by experts or native speakers of the named languages, the phonetic transcription task can be performed by both language learners and interested readers, who are unfamiliar with the languages as such.

In light of the above, we propose an application that allows the users to improve their understanding of the Persian poetry and language by engaging them with interactive tasks while also gathering data correlating the corrections to our transliterated corpus with the linguistic proficiency of our participants. Albeit the evaluation of the pedagogic influence of our application exceeds the limitations of the current effort, we pose the following questions:

- 1) Can accurate corrections of a phonetically transcribed corpus be acquired through our application in the form of learnersourcing?
- 2) How can the complexity of specific task be measured to allow for a more structured learnersourcing experience?
- 3) What are the pedagogical benefits of this type of activity, in terms of improved vocabulary learning with the audio annotations?

2 RELATED WORK

The problem of text-to-speech alignment (phonetic alignment) is approached with several methods such as use of Hidden Markov Models for alignment of speech on text [7, 25] or the combination of text-to-speech

¹<https://www.mturk.com/>

²<https://git.informatik.uni-leipzig.de/dsr/hafiz-prototype>

systems and the Dynamic-Wrapping algorithm [18]. These attempts are mostly limited to the languages such as English, French, German, and Spanish. However, there are efforts to expand this field by adding more languages [11]. The rise of crowdsourcing provided the opportunity for researchers to create multilingual transcribed corpora linked to the audio annotations [4, 5], but also to include low-resource languages, to avoid the slow and cost intensive process of transcriptions by experts [9]. Although the studies proved that the crowdworkers can achieve high data accuracy [9, 29], to our knowledge the task is mostly running on platforms such as Amazon Mechanical Turk, where micro-payments are offered to workers for performing micro-tasks.

The newly introduced concept of learnersourcing [15] describes the use of input by larger groups of learners to improve the learning video contents and interfaces. Some related projects in this context that also pursue the same objective aim at creating learnersourced subgoals in the framework of how-to videos [27], while some others crowdsource the assessments of exercises [21] or provide personalized hints [10] and generate explanations [30] for learners supporting them to solve the problems. Furthermore, in the field of semantic annotations studies are conducted to understand the effectiveness of learnersourcing through the inter-contextual linking of Quranic texts [3]. Though, the most akin project to our study is CrowdClass by Lee et al., which allows the same concept of on-task learning as intended in our project [17]. CrowdClass integrates the learning of scientific concepts through the so-called scaffold learning with the task of image classification. In this study the performance of the citizen scientist is investigated and the potential effect of learning while working on classification of astronomical images similar to the one of GalaxyZoo ³ is evaluated.

Studies in the field of language learning show that the different types of annotation leverage the language acquisition [2], [6], and audio annotations are beneficial for vocabulary learning due to the fact that in this way the sound form is captured by the phonological memory [12]. The audio annotations play a meaningful role specially for the learners who are not familiar with the alphabets and phonemes of the language they learn [19], [16]. But yet, to our knowledge there is no project that attempts to disambiguate or correct phonetic transcriptions by non-native speakers and language learners enabling them not only to learn the language with the help of audio annotations but also provide support materials that can be useful for other learners.

3 RESEARCH METHOD

3.1 Research Design

In this project, we investigate the extent to which the users with no knowledge of Persian are able to correct the phonetic transcriptions while listening to audio recordings. For this purpose, we confront the participant with disambiguation/correction tasks of varying complexity. In each task, the participant is provided with a range of plausible IPA-based transcriptions and an input field to add a complementary, probably more fitting phonetic transcription. These tasks can be grouped into a) disambiguation tasks, where the correct transcription is given as an option, and b) completion tasks, in which the users are supposed to add only the short vowels and c) correction tasks, where the correct transcription is missing, and it ought to be detected and inserted manually by the users. From technical point of view, the complexity of tasks is classified,

³<http://zoo1.galaxyzoo.org/>

Alternatives for ترک	Dictionary
<p>Please select the correct transliteration according to your understanding, or provide your own one below.</p> <p><input checked="" type="radio"/> tæræke</p> <p><input type="radio"/> tærke</p> <p><input type="radio"/> tæræk</p> <p><input type="radio"/> torke</p> <p><input type="radio"/> tærk</p> <p><input type="radio"/> tork</p> <p><input type="radio"/> Own suggestion <input type="text"/></p>	<p>ترک</p>
<input type="button" value="Discard"/>	<input type="button" value="Save"/>

Fig. 1. User interface for choosing the correct phonetic transcription according to the audio.

depending on a) whether a ground truth, provided by a language expert exists b) whether a data-set with links between the original text and the possible phonetic transcriptions is available and c) whether all options are machine generated, while all the short vowels are missing and the correct phonetic transcription is unknown.

While the first type of task shall be used to evaluate the language capabilities and listening skill of the user, the second and third types represent the actual and proposed learnersourcing by leaving the finding of the correct phonetic form to the crowd of learners. Before working on these tasks, the users are presented with an introductory text explaining the participation process and the basic concept of the tasks. The complexity and classification of the specific tasks as described above is unknown to the user. To answer the first research question addressed above we conducted a pilot study that is described in detail in below sections.

3.2 Data Collection

For our pilot study, we designed an interface designed and provided the users with six lines of Persian poetry with the original text written in Arabic letters, the corresponding IPA-based phonetic transcription [1] of the text, and the audio recording by one native speaker reading the text. For each recording, a speech-to-text service is used to create a word-by-word alignment of the source text and audio recording through time-stamps. Depending on whether there are different options for the pronunciation of the words, we inserted intentionally possible but wrong pronunciation for some random words in the displayed transliterated text. The users were confronted with 20 disambiguation items (in form of disambiguation by clicking on the correct option among the multiple choice options) and 2 correction items (by typing the correct form in the corresponding field). For all 22 items, a ground truth was available.

To allow a deeper analysis of the data gathered, the users were asked to create a profile before solving the tasks. In the profile, the user's L1 and L2 languages could be entered as well as the information about the user's age, gender, education and nationality. Afterwards, we asked the users to listen to the audio recordings and for those words, which were part of the exercise, to choose the correct pronunciation by

clicking on the correct answer in the multiple choice pop-up window, as shown in Figure 1. While users listened to audio, the corresponding word and its phonetic transcription turned into red, and they were able to pause, repeat or start over the audio play. By solving the phonetic transcription task in the form of disambiguation or manual correction, the users generated a database with possible corrections that can be used for various purposes, such as the enhancement of machine transliteration accuracy as discussed in the context of transliterations between Persian and English in [13].

3.3 Modeling of Phonetic Distances

As outlined in 3.1, we provided participants with various options of IPA-based transcriptions to choose from, so that we can calculate the error rate for each item. A plain binary correctness grading would certainly underestimate the complexity of each task, as the difficulty and complexity of the task may depend on the distinctive features of the to-be-changed phones. For instance, two consonants /t/ and /s/ share more phonological features than /m/ and /ʒ/. In the first pair there are only two differences in Manner features, but the latter pair has differences in all Class, Place, Laryngeal and Manner features [14], which means that distinguishing between /t/ and /s/ shall be more challenging than between /m/ and /ʒ/. This demands the calculation of distance weightings, which enables us to explore the correlation between the average error rate for each item and the estimated correction effort, i.e. the distance between the displayed form and the correct answer, for that particular item.

Fontan et al. provide a multifaceted approach of calculating distances between phonological features, by calculating phonologically weighted Levenshtein’s distance (PWLD) [8]. Based on a mapping of phonetic features of the IPA letters by Katamba [14], we calculate distances for all phonetic transcription options for consonants and vowels, respectively. Since not all the spoken vowels are existence in Persian phonology [20], feature mappings for vowels has been adapted to Persian, as just a subset of the features described by Katamba is used in Persian language.

For calculating the distance measure d , we proceeded as follows: Having a ground truth phonetic transcription w and a given answer u , for each letter, the basic distance is equal to the number of different phonetic features p divided by the total number of phonetic features n [8]:

$$d(w, u) = \frac{p(w, u)}{n} \quad (1)$$

The total number of phonetic features for consonants and vowels is 15 and 4, respectively. However, a stark weighting of a single feature difference occurs when treating vowels, which would result in a distance of $\frac{1}{4} = 0.25$ due to just 4 phonetic features related to vowels, as opposed to the 15 features related to consonants, $\frac{1}{15} \approx 0.067$. Thus, we employed a simple quadratic easing for vowel distances, d_v :

$$d_v(w, u) = d^2(w, u) \quad (2)$$

This results in $(\frac{1}{4})^2 = 0.0625$ and increases squarely as more vowel-related features differ.

3.4 Technical Implementation

When approaching the realization of the research method, we were facing three technical tasks: First, the corpus of Persian poetry we conducted was lacking audio annotation alignments. Second, this corpus was

to be presented to users via a user interface during user studies. And third, collected user data and study results were to be stored and, subsequently, evaluated within a database.

To gather audio annotation alignment data, we utilized the Google Cloud Text-to-Speech⁴ service. By using this service with our prerecorded spoken audio, configured to recognize Persian language, word-level alignments were acquired quickly for all of our corpus within short time. As the data lacked some inaccuracies, however, manual adjustments were necessary.

As our initial motivation arose from the Perseus Project [24], and, especially, its Scaife Viewer⁵ application, we opted to build a JavaScript-based single-page application (SPA) for users to interact with. The user interface library Vue.js⁶ has been adopted for this work, as it promised possible interoperability with Scaife Viewer for incorporating the audio alignment functionality. Data storage is provided by an Apache CouchDB⁷ NoSQL database, which offers convenient interaction capabilities via its HTTP interface. The data analysis pipeline then converts the filtered and enriched data into the CSV-based tabular data for further post-processing and analysis.

4 PILOT STUDY RESULTS

As stated in previous sections, due to the novelty of this project idea we conducted a pilot study to evaluate the methodology of data collection and refine the research questions. Hence, in scaffolding learning the adjustment of the scaffolding is one of the most essential and challenging steps. The prototype application went live for one week, during which a total number of 31 volunteers participated in the experiment. Of these 31 submissions only 16 are valid: Some submissions either contained corrupt data, or participants stated their knowledge of the Persian language, which meant that they did not fit within the scope of the study. Among the remaining 16 participants, four provided us with no profile, five of them entered German as their native language, four English, two French, and one Russian. Finally, none of these considered participants stated any familiarity with Persian.

For the analysis we identified three independent variables for each item; two of them were interval variables namely word length and the PWLD between the displayed text and the correct answer, i.e. what the user hears in the audio. The other parameter was a nominal parameter, which revealed if the correct answer was among the to-be-selected options or not. We weighted each item with the number of times that an answer was given to the option, and we explored the effect of the named parameters on the weighted error rate for each item.

A multiple linear regression was calculated to predict error rate based on PWLD, word length and the existence of correct answers among the options. A significant regression equation was found ($F(3, 216) = 135.238, p < 0.001$), with an R^2 of 0.653. Items' predicted error rate is equal to $-0.425 + 3.918$ (PWLD) -0.142 (word length) $+2.545$ (existence of the correct answer), where distance is measured as outlined above, length increased with each added character and the existence of correct answer is coded as 1 = existence, 2 = non-existence. Item's error rate increased for 3.918 depending on the increase of distance. Error rate decreased .142 for one character more in the length of the word, and it was 2.545 higher, when the correct

⁴<https://cloud.google.com/speech-to-text/>

⁵<https://scaife.perseus.org/>

⁶<https://vuejs.org/>

⁷<https://couchdb.apache.org/>

answer was not among the presented options. PWLD and the existence of the correct answer were significant predictors of error rate, whereas the word length was not a significant factor.

5 DISCUSSION

While the application created to conduct this study is still in the early stages of development, both the functionality and the user interface have been improved based on the participants' feedback. As the data analyzed in this study has been gathered throughout the development process, these inconsistencies in the workflow of our application are also adversely affecting the reliability of our data. Furthermore, the presented data is the result of a pilot study with one group design. The objective of this pre-experimental study was finding the deficiencies in the research design which helps us to improve the design and adjust our scaffolding. As the results of the pilot study indicate, the highest accuracy is achieved, when the displayed text is the correct answer, but when the displayed text is not the correct option, and it has to be selected among the options of multiple choice test, the accuracy drops down to 57.7%. However, if we decrease the complexity of the items to the lowest possible, i.e. eliminating the items with more than two options, and also the items, where the correct answer shall be entered manually, we achieve an accuracy equal to 67.8%. Alone this, emphasizes the importance of the scaffolding, which needs to be improved in the further stages.

Our initial research design also, as it happened, limited the capacity of our data. While some of the results were gathered with staff familiar with the project supporting the users in person, some users took part online without any assistance beyond the introductory text. From the feedback gathered, the in-person support was specially beneficial to users unfamiliar with the Persian language and the IPA. Additionally, creating a profile was optional, as was the amount of tasks the users solved before submitting the results. This also negatively influenced the reliability of our data. In future research, both a profile with a minimal set of information and completing all tasks presented to the user will be mandatory. If we want to further allow online participation, we will have to indicate these contributions to distinct them from results of on-site participation. Adding an interactive tutorial introducing the workflow of our application and the IPA to the user will also help to reduce the need for in-person support and allow for an preliminary evaluation of the users relevant capabilities.

6 CONCLUSIONS & FUTURE WORK

While our application provides the users with interactive tasks that we presume to have an positive effect on their learning experience, we need to develop the pedagogical structure to further the beneficial impact on their language capabilities. In order to grade the user's results, we use a modified algorithm for phonologically weighted Levenshtein's distance, which assigns a complexity to each option of a tasks based on it's ground truth. The results of the pilot study indicate that this distance has a significant effect on the error rate of the items, however, the extent of this impact and further parameters are still unknown to us. Conclusively, by evaluating these parameters we aim at creating a rating system, allowing us to order the tasks by complexity and present users with tasks matching their actual language capabilities, thus helping to create a more structured learnersourcing experience. Alleviated by this transformation, further research on the operationality of learnersourcing for the procuration of corrections to phonetically transliterated corpora will be facilitated.

As initially stated, the evaluation of the effect of our application on the understanding of Persian language, poetry and culture exceeded the limitations of this pre-experimental study. To permit the classification of the application as learnersourcing, additional research needs to be done in the form of two experiments. One experiment focuses on the influence of this kind of problem solving on the language capabilities of the users as a possible starting point. By analyzing the differences between vocabulary learning through this learnersourcing application with a ‘common’ vocabulary learning technique like flashcards with audio annotations, we can discover the prospects and measures necessary to further transform the learning experience of the users to their benefit. In this way, we will obtain a control group for comparison, which leads to the increase of reliability and validity of the study.

This experiment shall also be combined with a second experiment, in which the performance of ‘regular’ language learners on solving the defined task will be compared with the performance of the users with no background knowledge in that particular language in terms of data accuracy. The experiment still requires an appropriate scaffolding, consisting of the already mentioned tutorial in combination with protocol for the in person support provided during participation. In this regard, we believe that familiarity of users with the IPA and also adding some out-of-test exercises (e.g., a tutorial) will enable the users to perform more accurately during the testing phase.

The application itself is at an early stage of development. While all tasks used in this study were created manually by an expert, the intended learnersourcing of corrections to a transliterated corpus requires the machine generation of multiple, plausible, distinct transcription based on phonetic transcriptions of the users. While we have the ground truth for the initial evaluation of the users available, a rating system will be integrated into our application, so that the performance of the users can be voted where the task is not limited to disambiguation anymore and the transcription is required and no ground truth is available.

ACKNOWLEDGMENTS

We would like to thank all the individuals who shared their time for participating in the preliminary study described in this article.

REFERENCES

- [1] International Phonetic Association (Ed.). 1999. *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge University Press.
- [2] Lee B. Abraham. 2008. Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning* 21, 3 (2008), 199–226. <https://doi.org/10.1080/09588220802090246> arXiv:<https://doi.org/10.1080/09588220802090246>
- [3] Amna Basharat. 2016. Learnersourcing Thematic and Inter-Contextual Annotations from Islamic Texts. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 92–97. <https://doi.org/10.1145/2851581.2890386>
- [4] Andrew Caines, Christian Bentz, Calbert Graham, Tim Polzehl, and Paula Buttery. 2016. Crowdsourcing a multilingual speech corpus: recording, transcription and annotation of the CROWDED CORPUS. *Proceedings of LREC, Portoroz, Slovenia*.
- [5] Chris Callison-Burch and Mark Dredze. 2010. Creating Speech and Language Data with Amazon’s Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk (CSLDAMT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1–12. <http://dl.acm.org/citation.cfm?id=1866696.1866697>
- [6] Monica S. Cárdenas-Claros and Paul A. Gruba. 2009. Help Options in CALL: A Systematic Review. *CALICO Journal* 27, 1 (2009), 69–90. <http://www.jstor.org/stable/calicojournal.27.1.69>

- [7] Bert Van Coile, Luc Van Tichelen, Annemie Vorstermans, Jin Woo Jang, and M. Staessen. 1994. PROTRAN: a prosody transplantation tool for text-to-speech applications. In *ICSLP*.
- [8] Lionel Fontan, Isabelle Ferrané, Jérôme Farinas, Julien Pinquier, and Xavier Aumont. 2016. Using Phonologically Weighted Levenshtein Distances for the Prediction of Microscopic Intelligibility. In *Annual conference Interspeech (INTERSPEECH 2016)*. pp. 650–654. <https://hal.archives-ouvertes.fr/hal-01474904>
- [9] Hadrien Gelas, Solomon Teferra Abate, Laurent Besacier, and François Pellegrino. 2011. Quality Assessment of Crowdsourcing Transcriptions for African Languages. In *Twelfth Annual Conference of the International Speech Communication Association*.
- [10] Elena L. Glassmann, Aaron Lin, Carrie J. Cai, and Robert C. Miller. 2016. Learnersourcing Personalized Hints. In *CSCW'16*. Association for Computing Machinery, New York, NY, 1626–1636.
- [11] Jean-Philippe Goldman. 2011. EasyAlign: an automatic phonetic alignment tool under Praat. In *Twelfth Annual Conference of the International Speech Communication Association*.
- [12] Kirsten Hummel and Leif French. 2010. Phonological Memory and Implications for the Second Language Classroom. *The Canadian Modern Language Review / La revue canadienne des langues vivantes* 66 (03 2010), 371–391. <https://doi.org/10.3138/cmlr.66.3.371>
- [13] Sarvnaz Karimi. 2008. *Machine transliteration of proper names between English and Persian*. Ph.D. Dissertation. RMIT University. <https://researchbank.rmit.edu.au/view/rmit:161689>
- [14] Francis Katamba. 1989. *An Introduction to Phonology*. Longman.
- [15] Juho Kim. 2015. *Learnersourcing: Improving Learning with Collective Learner Activity*. Ph.D. Dissertation. MIT.
- [16] Batia Laufer and and Monica Hill. 2000. What lexical information do L2 learners select in a CALL dictionary and how does it affect word retention? *Language Learning & Technology* 3 (01 2000), 58–76.
- [17] Doris Jung-Lin Lee, Joanne Lo, Moonhyok Kim, and Eric Paulos. 2016. Crowdclass: Designing Classification-Based Citizen Science Learning Modules. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*. <https://www.aaai.org/ocs/index.php/HCOMP/HCOMP16/paper/view/14027>
- [18] Fabrice Malfèvre and Thierry Dutoit. 1997. High-quality speech synthesis for phonetic speech segmentation. In *EUROSPEECH*.
- [19] Dubois Michel and I Vial. 2001. Multimedia design: The effects of relating multimodal information. *Journal of Computer Assisted Learning* 16 (12 2001), 157 – 165. <https://doi.org/10.1046/j.1365-2729.2000.00127.x>
- [20] Corey Miller. 2012. Variation in Persian Vowel Systems. *Orientalia Suecana* 61 (01 2012), 156–169.
- [21] Piotr Mitros. 2015. Learnersourcing of Complex Assessments. In *Proceedings of the Second ACM Conference on Learning @ Scale*, Gregor Kiczales, Daniel M. Russell, and Beverly Woolf (Eds.). ACM Press, New York, New York, USA, 317–320. <https://doi.org/10.1145/2724660.2728683>
- [22] Joel Ross, Lilly Irani, M Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who Are the Crowdworkers? Shifting Demographics in Mechanical Turk. *Conference on Human Factors in Computing Systems - Proceedings*, 2863–2872. <https://doi.org/10.1145/1753846.1753873>
- [23] Mehrnoush Shamsfard. 2011. Challenges and open problems in Persian text processing. *Proceedings of LTC* 11 (2011).
- [24] David A. Smith, Jeffrey A. Rydberg-Cox, and Gregory R. Crane. 2000. The Perseus Project: a Digital Library for the Humanities. 15, 1 (April 2000), 15–25. <https://doi.org/10.1093/llc/15.1.15>
- [25] David Talkin and Colin W. Wightman. 1994. The aligner: text to speech alignment using Markov models and a pronunciation dictionary. In *SSW*.
- [26] Cornelia Veja, Julian Hocker, Christoph Schindler, and Stefanie Kollmann. 2018. Bridging Citizen Science and Open Educational Resource. In *Proceedings of the 14th International Symposium on Open Collaboration (OpenSym '18)*. ACM, New York, NY, USA, Article 15, 12 pages. <https://doi.org/10.1145/3233391.3233539>
- [27] Sarah Weir, Juho Kim, Krzysztof Z. Gajos, and Robert C. Miller. 2015. Learnersourcing Subgoal Labels for How-to Videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & #38; Social Computing (CSCW '15)*. ACM, New York, NY, USA, 405–416. <https://doi.org/10.1145/2675133.2675219>
- [28] Andrea Wiggins and Kevin Crowston. 2011. From Conservation to Crowdsourcing: A Typology of Citizen Science. In *Proceedings of the 2011 44th Hawaii International Conference on System Sciences (HICSS '11)*. IEEE Computer Society, Washington, DC, USA, 1–10. <https://doi.org/10.1109/HICSS.2011.207>
- [29] Jason Williams, I Dan Melamed, Tirso Alonso, Barbara Hollister, and Jay Wilpon. 2011. Crowd-sourcing for difficult transcription of speech. (12 2011). <https://doi.org/10.1109/ASRU.2011.6163988>
- [30] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z. Gajos, Walter S. Lasecki, and Neil Heffernan. 2016. AXIS: Generating Explanations at Scale with Learnersourcing and Machine Learning. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, Jeff Haywood, Vincent Aleven, Judy Kay, and Ido Roll (Eds.). ACM Press, New York, New York, USA, 379–388. <https://doi.org/10.1145/2876034.2876042>