Literature Review: SABCTXT

Creating a gold standard corpus of under-represented South African Languages

Mosamat Sabiha Shaikh
Computer Science
University of Cape Town
Cape Town Western Cape South
Africa
shkmos004@myuct.ac.za

ABSTRACT

South Africa is a linguistically diverse country with 11 official languages. Nine of the languages are termed low-resource languages due to there being not enough electronic documents available for them. The exceptions are English and Afrikaans. Natural language processing techniques require electronic documents to extract content for processing to enable computers to understand, interpret and generate spoken languages. The goal of the project, SABCTXT, is to investigate various existing techniques to determine if speech technology can be used to "listen" to news reports in chosen languages and produce accurate and appropriate transcripts. These transcripts will be compared against a gold standard corpus for quality assurance. The aim of this section of the project is to develop a gold standard corpus for some of the 11 official South African languages. Consequently, the literature review will discuss work done on corpus development for low-resource languages, existing corpora, crowdsourcing as a method for corpus development, and tools used for crowdsourcing.

KEYWORDS

Gold Standard Corpus, Crowdsourcing, Low-Resource Languages

1 INTRODUCTION

South Africa is a linguistically diverse country with 11 official languages [17]. Despite this, there is a glaring disparity in the availability of electronic textual resources in those languages (excluding English and Afrikaans) [18]. With the rapid and continuous growth of technology, it is becoming increasingly important for such resources to be available and accessible for Natural language processing.

The South African Broadcasting Cooperation (SABC), the government-run national broadcaster, is required to have programmes in all the South African languages thus making it a content-heavy audio source. The South African Centre for Digital Language Resources (SADILAR) has developed speech technology tools that are accessible to the public. This project as a whole aims to investigate various existing techniques to determine if language and speech recognition tools, most likely attained from SADILAR, can be used to produce accurate and appropriate transcripts of news reports in South African languages from SABC news reports. In the case of this project, the transcripts produced from the news channels and elsewhere will be compared to the gold standard corpus to validate it.

The purpose of this literature review is to shed light on work that has been done in corpus development in South Africa and elsewhere. Various techniques and methods to develop corpora will be explored with an emphasis on crowdsourcing. Notable corpora developed in South Africa for language include the NCHLT corpus for speech recognition and the Lwazi corpus [5][6].

2 CORPUS DEVELOPMENT

2.1 Corpora

A corpus is a collection of texts or text extracts that have been put together to be used as a sample of a language or language variety. It consists of texts that have been produced in 'natural contexts' (published books, ordinary conversation, letters, newspapers, lectures etc), which means it mirrors natural language [1]. Corpora can be used for a variety of purposes in the field of natural language processing. Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand textual and spoken words in much the same way human beings can [2]. Additionally, NLP uses computational techniques for learning, understanding, and producing human language content [19].

Language corpora can be classified as monolingual or parallel (also multilingual) [3]. A Monolingual corpus contains texts in one language only whereas a parallel corpus consists of two or more monolingual corpora. The corpora are the translations of each other. Trustworthy corpora are necessary for the

training and meaningful evaluation of algorithms [4]. These standard collections are called Gold Standard Corpora (GSC). Gold standard corpora, in the context of NLP, are manually annotated collections of text. They need to be developed carefully and systemically to use as dependable sources of information regarding languages. It is treated as the benchmark for assessing the precision of results obtained from natural language processing techniques. For high-quality gold standard corpora, multiple experts view the data independently and the inter-annotator agreement is computed to ensure quality.

2.2 South African Corpora

The development of linguistic resources for use in NLP is of utmost importance for the continued growth of research and development in the field, especially for low-resource languages [20]. South Africa remains behind the rest of the world in terms of electronic linguistic resources. In most non-OECD countries, there are not sufficient economic drivers for the creation of such resources through normal private-sector mechanisms, and the development of speech technology in the languages or dialects of those countries depends on public or philanthropic support for resource creation [5]. In South Africa, such support was provided by the national Department of Arts and Culture (DAC), which identified speech technology as an important tool in the development of the eleven official languages of the country. Notable corpora developed in South Africa for languages are the NCHLT broadband corpus for speech recognition and the Lwazi corpus, which is a telephone-based automatic speech recognition corpus.

The NCHLT is a collection of more than 50 hours of speech from approximately 200 speakers per language, in each of the eleven official languages of South Africa [5]. The corpus development process can be broken down into corpus design, prompt design, data collection, transcription, dictionary development, corpus selection and quality verification. The prompt design involved using electronic text data, but since most South African languages do not have such resources, Wikipedia was used as a source to generate prompts. There was an attempt to implement crowdsourcing, but it was deemed unsustainable. Data collection was a challenge due to issues regarding location, power and the Internet. These are issues that remain pertinent to the current situation in South Africa.

The three-year Lwazi project (2006-2009) produced the core tools and technologies required for the development of multilingual spoken dialogue systems in all eleven of South Africa's official languages and piloted the use of these technologies in government information service delivery [6]. The Lwazi ASR corpus consists of annotated speech data in 11 of the official languages of South Africa from approximately 200 speakers per language. Table 1 displays the size of the Lwazi corpus. Data was collected over the telephone.

Table1: Size of the Lwazi corpus

Language	#total minutes	#speech minutes	#distinct phones
Afrikaans	213	182	37
English (SA)	304	255	44
isiNdebele	564	465	46
Sepedi	394	301	45
Sesotho	387	313	44
Setswana	379	295	34
siSwati	603	479	39
Xitsonga	378	316	54
Tshivenda	354	286	38
isiXhosa	470	370	52
isiZulu	525	407	46

Limited availability of electronic data persists in many attempts to build corpora in South Africa. Thus, projects on building language corpora resort to obtaining data from South African government websites and documents [20]. This presents a limitation as government resources do not represent many aspects of spoken language.

A majority of South Africans are multilingual and hence code-switching occurs commonly and spontaneously [21]. Code-switching is the phenomenon of using more than one language within the same conversation or utterance [22]. Ewald van der Westhuizen and Thomas Niesler [21] introduce a speech corpus containing multilingual code-switching compiled from South African soap operas. The decision to use soap operas was because they are multilingual and showcase multilingual code-switching. The corpus was created by gathering digital video recordings of 626 South African soap opera episodes, and for each episode, extracting mono audio from the original source videos. The ELAN media annotation tool was used for segmentation purposes. ELAN is a linguistic annotation tool that was designed for the creation of text annotations for audio and video files of language use [23].

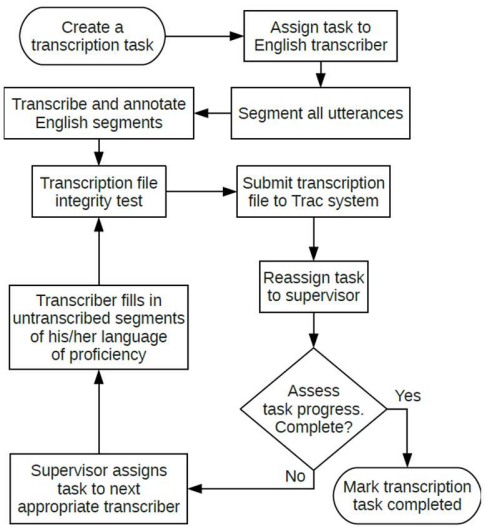


Figure 1: Flowchart of the transcription process

Figure 1 describes a flowchart depicting the transcription process for a single soap opera episode. The process of transcribing involves creating a task and assigning it to the main transcriber. The main transcriber is responsible for segmenting the utterances and transcribing them into English. Error checking occurs before submitting to the repository. Once submitted, the task is reassigned to the supervisor for overall progress assessment. If there are still incomplete segments, the task is given to another transcriber proficient in the relevant language. This continues until the supervisor confirms that the episode has been fully transcribed and marks the task as complete.

2.3 Other Low-Resource Corpora

The Bengali/Bangla language does not have much research done in the corpus development field, resulting in minimal corpora available for it. Nowshin *et al*. [7] propose a method of using crowdsourcing to develop a parallel corpus. A dataset of Bangla sentences and their corresponding English translations was collected through the means of crowdsourcing. The Bangla portion of the corpus consists mostly of simple and brief sentences with changes in tense, person, and sentence structure. The sentences included a variety of types, including imperative, negative, conditional, interrogative, assertive, and interrogative. The reason for the variation was that changes in sentence structure and results in translations provided could be observed. Textbooks on English grammar used in schools and English-to-Bangla translation books were the basis of the formation of the data. This limits the study to not being representative of the colloquial language encountered in the real world.

Salam *et al*. [8] attempt to build a balanced language corpus of the Bangla language at a national scale. The first phase sees the building of a Bangla Monolingual Corpus and the second and third phase consists of developing a multilingual parallel corpus. Due to its large scale, the building of the monolingual corpus (phase one) is focused on.

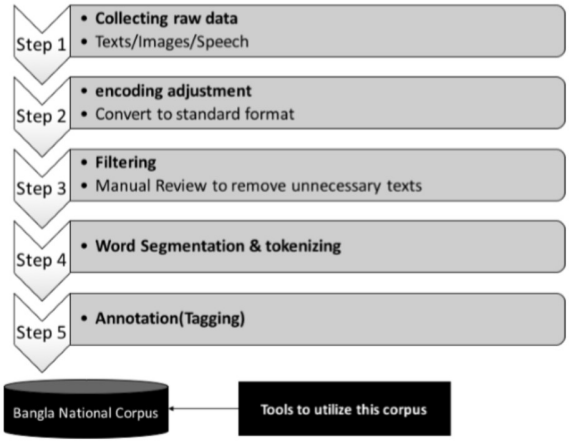


Figure 2. The development process of the Bangla corpus

Figure 2 depicts the steps of the process of building the monolingual corpus, which included:

Collecting Raw Data: Methods proposed were OCR, Web-crawling, typewriting, and using existing electronic text.

Encoding Adjustment: All the collected texts are converted to UTF-8 format using an encoding adjusting tool.

Filtering: Collected text was filtered for unwanted, unrecognized, foreign language, misspelt words, and garbage characters.

Word segmentation and tokenizing: Segmenting running text into words and sentences.

Annotation (Tagging): CoNLL-U, a standard format for annotation [24], was used to encode annotations in plain text files with three types of lines - Word lines, Blank lines, and Comment lines.

Hughes *et al*. [9] propose a system for quickly and cheaply building transcribed speech corpora containing utterances from many speakers in a variety of acoustic conditions. It uses a client-server system where the client is an Android device application written in Java that fetches textual prompts from the server and records the speaker's voice translation of the prompt. Figure 3 shows the client implementation of the system.

1 http://workshops.elda.org/lislr2010/sites/lislr2010/IMG/pdf/W2-AddaMariani-Presentation.pdf

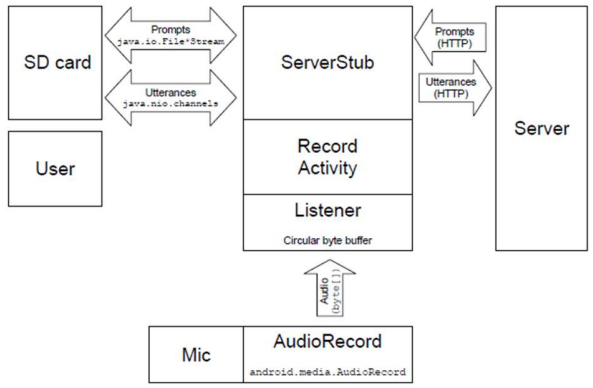


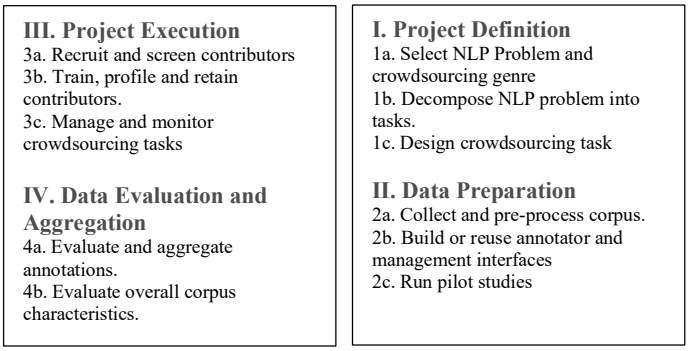
Figure 3: Client implementation

The server uses Web queries as prompts, which may not accurately represent the distribution of spoken language in the target environment. Furthermore, foreign words included in Web queries may be mispronounced or skipped by speakers, leading to inaccuracies in the speech data collected. The project aimed to provide an efficient, quick, and cheap method of building corpora in a variety of languages.

3 CROWDSOURCING

Crowdsourcing is the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers [10]. It is an emerging collaborative approach that can be used for the acquisition of annotated corpora and a wide range of other linguistic resources [11]. It has been used for speech transcription [16], system evaluation [25], read speech acquisition [26], search relevance[27], translation[28] and paraphrase generation[29,30]. Sabou *et al.* [11] break down the process of crowdsourcing into 4 main stages that are summarised in figure 4. The stages were deemed as best practice guidelines.

Figure 4: Crowdsourcing process



Project definition: An appropriate crowdsourcing genre has to be chosen followed by decomposing the NLP problem into simple tasks that can be completed by non-experts. Tasks should be simple and intuitive. When and how many contributors will be rewarded should be determined. The length of the text to be annotated needs to be kept reasonably short, without compromising accuracy.

Data preparation: Data may need to undergo annotation or filter to remove undesirable content. Interfaces need to be designed in a way that reduces cheating in crowdsourcing tasks.

Project execution: This is the main phase of the process**.** It consists of 3 kinds of tasks: task workflow and management, contributor management and quality control. It is highlighted that a core challenge for all crowdsourcing approaches is motivating contributors to participate.

Data evaluation and aggregation: The input of the contributors should be assessed. The goal is to make acquisition tasks reproducible, and scalable, and to ensure good corpus quality.

3.1 Custom Tools

Nowshin *et al.* [7] collect a dataset of Bangla sentences and their corresponding English translations through the means of crowdsourcing. A Web interface was used to collect English translations of Bangla sentences, randomly chosen from a Bangla corpus, from a group of undergraduate university students who were proficient in both Bangla and English.

The Android device application used by Hughes *et al*. [9] was deemed easy to use with a comparatively inexpensive setup. This made it possible for a large number of unskilled users to collect speech data in parallel. University students were tasked with recruiting speakers due to their proficiency in technology.

Packham and Suleman [31] developed a custom crowdsourcing system that employed gamification to gather multilingual content for building language corpora for low-resource languages. Gamification is an umbrella term for the use of video game elements (as opposed to full-fledged games) to improve user experience and user engagement in non-game services and applications [32]. Having a reward system is a common aspect of gamification. Thus, the custom crowdsourcing system had a scoring mechanism that was designed to have one-to-one mapping to money earned. 4 experiments regarding payment were conducted, and it was concluded that monetary payments played a larger role in motivating participants than gamification in tasks with strong intrinsic motivation.

3.1 Crowdsourcing Platforms

YangWang *et al*. [12] used Microsoft's Universal Human Relevance System (UHRS) crowdsourcing platform crowdsourcing methods to acquire language corpora for use in natural language processing systems. UHRS is a crowdsourcing platform that supports data labelling for various Artificial Intelligence application scenarios. Vendor partners provide and enable connections with people—who are referred to as 'judges'—to provide data labelling at scale. All UHRS judges are under NDA so data is always secure[13]. Constraints of not being able to set a maximum number of tasks assigned to judges were identified. This resulted in repeated sentences as a caveat.

Amazon's Mechanical Turk (MTurk), similar to UHRS, is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually [14]. Post *et al.* [15] built parallel corpora for 6 low-resource and under-studied languages with efficacy. Marge *et al*. [16] investigated whether the MTurk service can be used as a reliable method for the transcription of spoken language data. It concluded that it can indeed be used to accurately transcribe spoken language. Gelas *et al.* [33] evaluated the quality of speech transcription obtained via crowdsourcing and concluded that it is possible to acquire quality transcriptions from the crowd for under-resourced languages using MTurk. Legal and ethical[1] surrounding this tool exists and hence a set of guidelines of good conduct while using MTurk for research:

- Systematically explain "who we are", "what we are doing" and "why" in HITs descriptions (as done traditionally for data collection)
- Make the data obtained available for free to the community.
- Set a reasonable payment so that the hourly rate is decent.
- Filter contributors by country of residence to avoid those who consider MTurk as their major source of funding.

4 DISCUSSION

A recurring theme in many of the papers is the costs associated with building corpora. Wissler et al. [4] highlight that it is costly, time-consuming, and laborious to construct a gold standard corpus. Crowdsourcing, expert review and active selection schemes were suggested as means to significantly lower costs and retain the quality of corpora. Similarly, Barnard *et al.* [5] mention the need for economic drivers to facilitate the creation of resources. The NCHLT corpus was designed and developed through support provided by the Department of Arts and Culture. Cost-effectiveness was an important consideration during the design of the Lwazi corpus [6]. It was designed to be as small as possible while retaining usability. The system developed by Hughes *et al.* [9] for building corpora quickly and cost-effectively was tested in a variety of languages yielding positive results.

Lack of resource availability is an issue for low-resource languages. The creation of speech technology and consequently corpora is strongly tied to resource collection [5]. Not only do the resources have to be available but also representative of both spoken and written forms of the language being represented. Salam *et al*[8] attempt to build a representative and balanced corpora was not completed due to limited time and resources. When developing text resources for 10 South African languages, the biggest constraint in terms of attaining data was that all data collected would be made available as open-source resources [20]. Data providers were apprehensive because of the release of data with no limitations. Consequently, most of the data was sourced from South African government websites. This posed a limitation as government documents are not representative of the languages. Niesler *et al.* [21] designed a corpus to accommodate multilingual code-switching. This was a first for South Africa.

As mentioned previously, crowdsourcing for corpus development is cost-effective and retains quality. Nowshin *et al* [7] note that there is an advantage of obtaining insights into human behaviour during crowdsourcing. However, there is the caveat of being unable to control important variables such as the number of tasks per worker [12]. Packham and Suleman [31] found that gamification by itself does not yield increased motivation and engagement for seemingly important crowdsourcing tasks. There have to be financial incentives.

Tools for crowdsourcing exist such as MTurk and UHRS. While there were constraints to the number of tasks that could be set in UHRS, through experiments and analysis it was determined that it can gather high responses from workers with low latency [11]. On the other hand, Post *et al.* [15] note that there is a high variance in the quality of the translations obtained on MTurk. Additional tasks had to be designed for quality assurance. However, experiments conducted by Marge et al. [16] and Gelas *et al.* [33] conclude that MTurk can be used to accurately transcribe spoken language. Finally, legal and ethical issues surrounding crowdsourcing need to be considered [11]. Namely: how to properly acknowledge contributions, how to ensure contributors' privacy and well-being and how to deal with consent and licensing issues.

5 SUMMARY

The literature review discusses corpus development for low-resource languages around the world and in South Africa. The process of building corpora is outlined starting from data collection to annotation and quality control. Notable corpora built in South Africa include the Lwazi corpus and the NCHLT corpus. Crowdsourcing is discussed as an approach to

1 http://workshops.elda.org/lislr2010/sites/lislr2010/IMG/pdf/W2-AddaMariani-Presentation.pdf

building corpora. Best practices for crowdsourcing can be broken down into project definition, data preparation, project execution, and data evaluation and aggregation. Finally, Microsoft's Universal Human Relevance System (UHRS) and Amazon's Mechanical Turk (MTurk) are introduced. It is possible to collect quality transcriptions for under-resourced languages using these crowdsourcing tools.

REFERENCES

[1] Corpora - English Language: a short guide to online resources. Retrieved March 23, 2023 from https://libguides.bodleian.ox.ac.uk/english-language/Corpora

[2] What is Natural Language Processing? | IBM. Retrieved March 13, 2023 from https://www.ibm.com/topics/natural-language-processing

[3] Corpus types: monolingual, parallel, multilingual… | Sketch Engine. *Sketch Engine*. Retrieved March 13, 2023 from https://www.sketchengine.eu/corpora-and-languages/corpus-types/

[4] Wissler, Lars & Almashraee, Mohammed & Monett, Dagmar & Paschke, Adrian. (2014). The Gold Standard in Corpus Annotation. 10.13140/2.1.4316.3523.

[5] E. Barnard, M. H. Davel, C. J. van Heerden, F. De Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proceedings of SLTU*, 2014, pp. 194–200.

[6] E. Barnard, M. Davel, and C. van Heerden, "ASR corpus design for resource-scarce languages," in *Proc. Interspeech*, Brighton, UK, Sept. 2009, pp. 2847–2850.

[7] Nowshin, N., Ritu, Z. S., & Ismail, S. (2018, December). A Crowd-Source Based Corpus on Bangla to English Translation. In 2018 21st International Conference of Computer and Information Technology (ICCIT) (pp. 1-5). IEEE.

[8] K. M. A. Salam, M. Rahman, and M. M. S. Khan, "Developing the bangladeshi national corpus-a balanced and representative bangla corpus," in *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pp. 1–6, IEEE, 2019.

[9] Thad Hughes, Kaisuke Nakajima, Linne Ha, Atul Vasu, Pedro Moreno, and Mile LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *Proceedings of Interspeech*, 2010

[10] Crowdsourcing Definition & Meaning - Merriam-Webster. *Merriam-Webster*. Retrieved March 13, 2023 from https://www.merriamwebster.com/dictionary/crowdsourcing

[11] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl, "Corpus annotation through crowdsourcing: Towards best practice guidelines," in *Proc 9th Int. Conf. Lang. Resour. Eval.*, 2014, pp. 859–866.

[12] William YangWang, Dan Bohus, Ece Kamar, and Eric Horvitz. 2012. Crowdsourcing the acquisition of natural language corpora: Methods and observations. In 2012 IEEE Spoken Language Technology Workshop (SLT). IEEE, Piscataway, New Jersey, US, 73–78. https://doi.org/10.1109/SLT.2012.6424200

[13] UHRS. *UHRS*. Retrieved March 13, 2023 from https://prod.uhrs.playmsn.com/

[14] Amazon Mechanical Turk. *Amazon Mechanical Turk*. Retrieved March 13, 2023 from https://www.mturk.com/

[15] Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the 7th Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 401–409.

[16] M Marge, M. Banerjee, S. & Rudnicky, A. I. (2010). Using the Amazon Mechanical Turk for transcription of spoken language. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 5270– 5273). Retrieved from http://dblp.uni-trier.de/db/conf/icassp/icassp2010.html

[17] Jan Alewyn Nel, Velichko H Valchev, Sebastiaan Rothmann, Fons JR Vijver, Deon Meiring, and Gideon P Bruin. 2012. Exploring the personality structure in the 11 languages of South Africa. 80, 4 (2012), 915–948. *Journal of Personality*

[18] Shigeaki Kodama. 2008. Languages on the Asian and African Domains. In *Proceedings of the International Symposium* on CDG. 77–82.

[19] Julia Hirschberg and Christopher D. Manning. 2015. Advances in natural language processing. *Science* 349, 6245 (2015), 261-266. DOI: https://doi.org/10.1126/science.aaa8685

[20] R. Eiselen and M. J. Puttkammer, "Developing Text Resources for Ten South African Languages," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, May 2014, pp. 3698–3703.

[21] Thomas Niesler et al. 2018. A frst South African corpus of multilingual code-switched soap opera speech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[22] Ondene Van Dulm. 2009. English–Afrikaans intrasentential code switching: Testing a feature checking

account. *Bilingualism: Language and Cognition* 12, 2 (2009), 193-212. DOI: https://doi.org/10.1017/s1366728909004039

[23] H. Brugman and A. Russel. Annotating multi-media / multimodal resources with ELAN. *Proceedings of the 4th International Conference on Language Resources and Language Evaluatio*n (LREC 2004), pages 2065–2068, 2004.

[24] L. B. Villa, Udeasy: a tool for querying treebanks in conll-u format, in: Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-10), 2022, pp.16–19.

[25] Yang, Z., B. Li, Y. Zhu, I. King, G. Levow, and H.M. Meng, Collection of user judgments on spoken dialog system with crowdsourcing, In *Proc. of SLT*, 2010.

[26] Lane, I., M. Eck, K. Rottmann and A. Waibel, Tools for Collecting Speech Corpora via Mechanical-Turk., In *Proc. Of Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[27] Alonso, O., D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation, In *Proc. of SIGIR Forum 42*, 2008.

[28] Zaidan, O., C. Callison-Burch, Crowdsourcing Translation: Professional Quality from Non-Professionals, In *Proc. of ACL*, 2011.

[29] Burrows, S., M. Potthast, and B. Stein. Paraphrase Acquisition via Crowdsourcing and Machine Learning. In *ACM TIST* (to appear), 2012.

[30] Dolan, W. B., C. Brockett. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proc. of The Third International Workshop on Paraphrasing*, 2005.

[31] Sean Packham and Hussein Suleman. 2015. Crowdsourcing a Text Corpus is not a Game. In *Digital Libraries: Providing Quality Information: 17th International Conference on Asia-Pacific Digital Libraries, ICADL 2015, Seoul, Korea, December 9-12, 2015. Proceedings*, Robert B. Allen, Jane Hunter, and Marcia L. Zeng (Eds.). Springer International Publishing, Cham, 225–234. DOI: http://dx.doi.org/10.1007/978-3-319-27974-9_23

[32] Deterding, S., Sicart, M., Nacke, L., O'Hara, K., and Dixon, D. Gamification: Using game-design elements in nongaming contexts. *Proc. CHI EA '11*, ACM Press (2011), 2425-2428.

[33] Gelas H., Abate S.T., Besacier L. and Pellegrino F. (2011). Evaluation of crowdsourcing transcriptions for African languages. In *HLTD 2011*.

1 http://workshops.elda.org/lislr2010/sites/lislr2010/IMG/pdf/W2-AddaMariani-Presentation.pdf