Studies in African Linguistics
Volume 52 Supplement 13, 2023.

Reflections on the development of corpora for Zimbabwe's understudied languages

Emmanuel Chabata
University of Zimbabwe

Linguistic corpora are one of the primary research tools in modern-day linguistics. The centrality of corpora derives from the philosophy that data from them is more accurate, observable, objective, reliable, and verifiable. However, very little has been done to develop corpora in understudied languages. Yet compiling readily available corpora is principally important for these languages since most researchers have restricted physical access to them given that most of them are in remote areas. This article examines issues of corpus designing, compilation, and querying and is a call for the development of corpora in Zimbabwe's understudied languages. Taking a cue from some of the challenges encountered in the development of Shona and Ndebele language corpora, the article focuses on issues that need special consideration when developing corpora in these languages. Some such issues relate to the languages' level of development, the scarcity of written and electronic materials in them as well as the sociolinguistic context in which they are found. An argument is made that corpora should be developed in these languages so that they become an important footing upon which the development of other linguistic resources can be anchored.

Keywords: Corpus, linguistic data, understudied language, Zimbabwe

1. Introduction

Linguistic corpora are one of the primary research tools in modern-day linguistics. They permit researchers from all occupations who are stationed anywhere across the world to study raw language data of any language rather than its interpretations by earlier linguists in grammar books or journal articles. In fact, reliance on corpora in the study of language structure and use has become indispensable in contemporary linguistic applications (see, for example, De-Schyriver and Prinsloo, 2000; Kennedy, 1998 and McEnery and Wilson, 2001, Midrigan-Ciochina, et al. 2020). The centrality of corpora as sources of linguistic data derives from the philosophy that data from them is more accurate, observable, objective, reliable, verifiable, and easy to access. They provide empirical data, which supports accountability and falsification. Leech (1992) proposes falsification as a desirable trait of linguistic models based on open corpora. As noted in McEnery and Wilson (2001:1), studying language using data from a well-constructed corpus is studying language "based on examples of 'real life' language use". Dependence on corpus data is thus a way of circumventing the known weaknesses of the traditional sources of linguistic evidence such as intuition and introspection. For example, because the researcher's mind is the only source of evidence, the data from intuition and introspection is generally incomplete, non-observable, unverifiable, subjective, and idiosyncratic. Corpora become even more important in contemporary linguistic research, which Walther and Sagot (2017:89) argue "relies more and more heavily on the exploration of statistical patterns in language".

In reference to the importance of corpora in the study of African languages, De-Schyriver and Prinsloo (2000:1) contend that; "... if African linguistics is to take its rightful place in the new millennium, the active compilation, querying and application of corpora should therefore become an absolute priority". This argument was made after an observation that most African languages do not have this important language resource. The observation made by De-Schyriver and Prinsloo more than two decades ago can still be made today because nothing much has changed since then. The few languages that have corpora are mainly national languages that have always been prioritized for development since the beginning of language studies. The same languages continue to get preferential treatment even today when the new consciousness directs that all languages should be treated equitably.

There are no corpora developed in understudied or under-documented languages. This is because these languages are marginalized from all forms of development. Their lack of development is owing to a variety of reasons ranging from lack of funding and other resources to logistical problems, low speaker numbers, and low vitality. Yet compiling readily available corpora is principally important for marginalized languages since most researchers have restricted physical access to them given that most of them are located in remote areas of their respective countries.

This article is an argument for the development of corpora in Zimbabwe's once-marginalised languages, namely Chewa, Chibarwe, Kalanga, Koisan, Nambya, Ndau, Shangani, sign language, Sotho, Tonga, Tswana, Venda and Xhosa. For the purposes of this article, we prefer to call these languages 'understudied languages' owing to the fact that they are materially under-resourced, and generally, there is no continuous text production in them. The languages also lack a recognisable literary tradition just as they lack any significant digital presence. For the reason that written resources are scarce in such languages, Maxwell and Hughes (2006: 30) prefer to call them 'lower-density languages'. This term is used in contrast to languages such as English, Chinese, and other European languages that the two scholars refer to as 'high density' owing to the abundance of resources in them. As noted in Szymanski (2011:8), under-resourced languages may not necessarily be endangered or minority, although they might be.

The understudied languages that are the focus of this article are identified in the Zimbabwean Constitution as the country's officially recognized languages together with English, Ndebele, and Shona. Whilst English, arguably the world's biggest language in terms of use and influence, boasts some of the biggest and oldest corpora, Shona and Ndebele benefited from the African Languages Lexical (ALLEX) Project – African Languages Research Institute (ALRI) research programme that developed corpora in them. As the country's national languages, Shona and Ndebele have also been favoured ahead of the rest of the other indigenous languages in terms of general language development and in terms of the development of language resources. Compared to all the other Zimbabwe indigenous languages that have been marginalized for a long time, Shona and Ndebele can arguably be described as well-studied or well-resourced. They have relatively long literary traditions and have a variety of literature developed in them. Owing to the availability of this literature, the corpora in the two languages were constructed from both written and oral materials.

The article examines issues of corpus designing, compilation, and querying as they relate to understudied languages. Taking a cue from some of the challenges encountered in the development of Shona and Ndebele corpora, the article focuses on issues that need special consideration if corpora have to be developed in these languages. Some such issues relate to the level of development in these languages, the scarcity of written and electronic materials in them as well as the sociolinguistic context in which they are found. An argument is made that corpora should be developed in these languages so that they become an important footing upon which the development of other linguistic resources can be anchored. With corpora as sources of linguistic evidence, the documentation and description of these languages would become faster and more authentic as they would be based on instances of natural language use.

2. A standard corpus

Although the focus of this article is corpora for understudied languages, a brief discussion of what a standard corpus entails may be necessary if one has to appreciate the misgivings associated with corpora in under-resourced or under-documented languages. The term *corpus* has been in use even before the advent of computers to refer to language collections in a variety of forms. The collections varied from texts written in long hand to spoken language data recorded on tape or disc. In terms of size, the corpora were generally small to the extent that any collection of more than one language text that is used as a source of evidence in linguistic analysis was regarded as a corpus. However, in what they refer to as modern linguistics, McEnery and Wilson (2001:32) define a linguistic corpus as, "…a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration". From this description, three characteristics emerge as the core of falsifying corpus-based linguistic analysis as a plausible methodology: size, machine-readability, and representativeness.

Unlike the earlier conception where the term corpus referred to any collection of texts irrespective of the form in which they are, the characterization by McEnery and Wilson (2001) points to electronically accessible texts. This entails that even though corpus materials may be collected in other forms such as speech and printed texts, eventually the materials should be turned into machine-readable form. In this form, language-engineering tools can be applied to make it easier and faster to search, sort, and manipulate the data. That is, the researchers can store, search through, and organise huge amounts of data by the click of a mouse. Statistical information also becomes easier to get from the language collection.

It should be noted that earlier corpora tended to be small. This was probably caused by limitedness in terms of sources for data collection and storage facilities as well as difficulties associated with manual data analysis. However, with the current setup where the internet has become a huge source of corpus data, it has become relatively easy to construct corpora that run into billions of tokens. Depending on the intended use(s) of the corpus, one can decide to build a finite-sized or a monitor corpus. A monitor corpus is described in McEnery and Wilson (2001:30) as one in which, "Texts are constantly being added…so that it gets bigger and bigger as more samples are added". For a finite-sized corpus "…the research plan will set out in detail how the language variety is to be sampled, and how many samples of how many words are to be collected so that the pre-defined grand total is arrived at" (McEnery and Wilson, 2001:31). If the researchers are interested in quantitative analysis, finite-sized corpora will be ideal while monitor corpora are more suitable for a variety of qualitative research methodologies useful in studies such as lexicography, linguistic change and terminology development. Although the size of a corpus, together with other aspects such as representativeness, does affect its validity and reliability, it should be noted that the issue of size alone might not be fundamental in determining the usefulness of a corpus. This is because "…any corpus, however big, can never be more than a minuscule sample of all the speech or writing produced or received by all the users of a major language on even a single day" Kennedy (1998:66). We should also hasten to note that huge corpora are only achievable in major languages that have rich sources of texts such as the internet.

Of all the characteristics that define a prototypical corpus, representativeness is the most elusive. This concept requires that a corpus should be filled with a wide range of samples from different authors or speakers and genres which, when taken together, may be considered to 'average out' and provide a reasonably accurate picture of the entire language population the researcher is interested in (McEnery and Wilson, 2001). Typically, such a corpus should be balanced in that it should represent variation across all types of data sources and cover all types of corpus sources relevant to a group of people's linguistic practices (Cox, 2010). For example, a representative corpus is expected to contain all types of oral and written texts present in the language, that is, various genres of fiction (oral, literary, and scientific), academic, journalistic, business, dialectal, and sociolectal texts, etc. The corpus should also be balanced in that the sizes of data subsamples are proportional to the proportions of the speakers, registers, varieties, etc. in the total population the corpus is meant to represent (Vinogradov, 2016). To construct such a corpus, the corpus designers would need to come up with clear strategies for text selection; that is, texts should be selected in a way that brings balance to the total language population. To be seriously considered in terms of balance are demographic, social, geographical, and textual issues. A representative corpus, as a language sample, would enable the researcher to generalize the findings from that corpus to the total text population represented. In other words, lack of representativeness tends to limit the research questions that may be asked as well as the degree to which generalizations derived from such a corpus can be applied to other situations (Biber, Conrad, and Reppen, 1998:246).

Having noted the key characteristics that define a prototypical corpus, one wonders whether these are achievable in the development of corpora for all languages. One also doubts whether the demand for these would not stand in the way of constructing corpora in less-resourced languages where issues of representativeness and a huge corpus size, for example, may be difficult, if not impossible, to achieve. The question would be whether corpora that do not meet some of these defining characteristics should be accepted as sources of authentic data. In the succeeding section, we characterize the kinds of corpora that are possible or achievable in languages that are understudied, under-documented, and under-resourced.

3. Corpora of understudied languages

We have already indicated that linguistic corpora are rare in understudied languages. Their non-availability is owing to a number of factors some of which include (a) the general marginalization that these languages have endured for a long time, (b) lack of attention by language researchers and other stakeholders due to their low vitality, (c) lack of funding or commercial interest, amongst others. Yet corpora are imperative in these languages for a number of reasons.

For starters, most understudied languages are either poorly documented or not documented at all. Some of them are even threatened with extinction. In this case, corpus construction would serve as a form of language documentation and language preservation. Thus, although it is not part of their history, corpus development in these languages may be conceived of as a strategy in response to language endangerment. The aim would be to construct permanent, reusable, and multifunctional collections of primary linguistic data that can be used in a wide range of linguistic and socio-cultural applications. For example, well-built corpora in these languages would be useful in the production of important language reference works such as dictionaries, descriptive grammars, vocabulary lists, etc.

Besides their importance as sources of data in the production of reference works, corpora in understudied languages are an important resource in the development of language engineering applications that will in turn enhance linguistic analyses. Basic language engineering products such as spell checkers, morphological analysers and syntactic analysers are easier to develop when standardized corpus data is available. This is probably why McEnery, Baker, and Burnard (2000) note that corpus data is the *sine qua non* of many language-engineering applications. Without corpus data, the ability of language engineers to generate tools or systems for use within a language is seriously reduced. The need for technological tools or systems for linguistic analysis in these languages would therefore be one of the important reasons why researchers and other stakeholders should prioritise corpus development. Commenting on the benefits of corpus linguistic data in the case of underprivileged languages, McEnery and Ostler (2000:417) note that,

> …in collecting corpus data for endangered and minority languages, researchers are not only paving the way towards better descriptions of those languages, they may also be allowing work to begin on [a] range of language technology application… Corpora are multifunctional resources, and their potential for reusability is high.

Corpus construction in understudied languages differs in some fundamental ways from that of well-studied or well-documented languages. The discrepancy is mainly caused by differences in linguistic infrastructure between the two language categories. For example, whilst there is an abundance of a wide variety of written texts (both printed and electronic) in well-studied languages, such materials are generally lacking in understudied languages. For that reason, it is easier to build huge and representative corpora consisting of written and spoken texts in well-studied languages compared to what is possible in understudied languages. With the paucity of written texts in understudied languages, corpus collections are mostly based on audio recordings of spoken texts that are later transcribed and annotated. Compared to corpus construction in well-studied languages where different kinds of written texts from a variety of sources are readily available in electronic form, recording and transcribing oral data is a significantly long, slow, and expensive process of data collection and processing. Because of that, the possible corpora in these languages are generally small. Also because of the scarcity of written materials in whatever form, no meaningful text selection is possible in these languages. Vinogradov (2016:135) notes that "The developers of small corpora of understudied languages cannot afford the luxury of rejecting available texts, because such languages are normally under-resourced". Without text selection, it is not possible to talk about representativeness or balance in the construction of corpora in these languages.

With the issues of representativeness (quality) and sample size (quantity) both tilting against the corpora of understudied languages, a question can be raised on the utility of such language collections. Whilst the instinctive response would be a temptation to discard the corpora as inadequate and unreliable, a thoughtful reaction would consider two fundamental issues; (a) the feasibility of constructing huge and balanced corpora given the weakened linguistic infrastructure in which the languages are found, and (b) the goal(s) of constructing corpora in these languages.

Our considered view is that insisting on building the so-called standard corpora in understudied languages would only ruin the chances of having any language collections in them. On the question of balance in corpus development, for example, we tend to concur with Sinclair (2005) who notes that although it must be used to guide the design of a corpus and the selection of its components, the notions of balance and representativeness are not precisely definable and attainable goals. Corpus construction in understudied languages may not be stalled by unachievable goals. Some deviation from the prototype must not disqualify a language collection from being a language corpus if the collection can still fulfil a wide range of language documentation and language analysis objectives. In fact, the collection should be accepted as a useful research resource as long as it fulfils the primary objective of linguistic corpora described in Vinogradov (2016:136) as "…to help linguists find and explore sentences (occurrences) in texts [in a particular language] that meet specific search criteria". In other words, a corpus should be considered good and useful as long as it supports the researchers in doing what they wish to do. Whilst huge and well-constructed corpora would be desirable as forms of language documentation and preservation and as bases for the development of language engineering tools and systems, it is gratifying to note that the relatively small and poorer quality corpora possible in understudied languages can still serve these purposes, which can, in turn, incentivize the development of corpora typical of those developed in well-studied languages. If they are well annotated, such corpora can also be useful sources of data in language enterprises such as grammatical descriptions and the editing of dictionaries.

4. Corpus development in Zimbabwe's understudied languages – matters for reflection

In this section, we look at a few issues that we feel need reflection in the construction of corpora in Zimbabwe's understudied languages already identified as Chewa, Chibarwe, Kalanga, Khoisan, Nambya, Ndau, Shangani, sign language, Sotho, Tonga, Tswana, Venda, and Xhosa. Although these languages are under-resourced, the goal should be to produce reusable collections of primary linguistic data that Stubbs (2001:66) calls 'first-order data'. The aim should be to build permanent language resources that support both empirically grounded linguistic research and efforts at addressing issues of language endangerment. In this case, the starting point could be to consider corpus development as a form of language documentation, a means through which the corpus builders can preserve traces of respective communities' cultural heritages through language. Priority could also be given to the construction of general (not specialised) corpora suitable for a diversity of analyses and acceptable to speech communities whose languages they represent.

The principles proposed for the development of such corpora converge with those of documentary linguistics whose key features are captured in Nathan (2009:102) after Himmelmann (2006:15) as follows:

- *focus on primary data* – consisting of collecting and analysing an array of primary language data which is also made available to a wide range of users
- *accountability* – access to primary data and representations of it in a way that makes for a more transparent evaluation of linguistic analyses
- *long-term preservation* – a focus on archiving to ensure that linguistic materials are available to a range of potential users into the distant future
- *interdisciplinary teams* – documentation requires input and expertise from a range of disciplines and is not restricted to linguists alone
- *involvement of the speech community* – collaboration with community members not only as consultants but also as co-researchers.

In trying to achieve these features, researchers and other stakeholders may need to seriously consider the following issues that are pertinent in the development of any language collection, but that need more thoughtfulness with regard to understudied languages.

4.1 Corpus planning. Careful planning to determine the prospective uses of the corpus, its target users as well as to identify data sources and the methods to be used in data collection should precede any process of corpus development. The planning process should revolve around the goal of producing a language collection that represents a wide variety of language styles so that in the final analysis the corpus can be said to be a microcosm of the concerned language in its current state. Just as in the construction of any corpus, the development of corpora in understudied languages should not only consider the features of the available linguistic material, for example, its representativeness in terms of spoken vis-à-vis written languages and social and regional dialects but should also be sensitive to issues of digital representation that include the selection of appropriate standards for text transcription and encoding.

Clear-headed planning around these issues is more critical when resources for corpus development are limited. The corpus planner should grapple with questions such as (a) which features should be given immediate attention in corpus construction, and which ones should be set aside as areas for future development? (b) Which subset of prospective corpus users and corpus uses should be prioritised for specific focus in the short-term development? and (c) What linguistic features should be selected for annotation and at what level of detail?

4.2 Data collection. Data collection activities are important in corpus development as they are the ones that determine corpus content. The collection and processing of data in any corpus construction activity should follow the standard procedure as much as possible; that is, the relevant research and methodologies are theoretically expected to be essentially similar in all languages. For example, when collecting data, the corpus developers should generally be guided by the need to adequately cover all possible linguistic contexts. However, language situations vary in ways that result in deviations from the so-called standard procedure. For example, unlike in well-studied languages where data can be obtained from a multitude of sources, most of which involve data already in electronic form, the situation is different for Zimbabwe’s understudied languages that do not enjoy long histories of writing and where some do not even have standard orthographies. The obtaining situation entails that there are very few written materials (in either printed or electronic form) for use in corpus construction in these languages. That leaves the spoken language as the sole important source of corpus data. That is, data for corpus development would mainly consist of transcriptions of audio-recorded data collected through recording of family and friends’ everyday conversations, recording radio programmes such as news bulletins in the respective languages, interviews, religious meetings and sermons, school lessons, recipes, procedures, instructions, commentaries, narratives, poetry, song, drama, prayer, laments, jokes, amongst many others.

With the spoken language being the chief source of corpus data, issues such as the requisite equipment for data collection, the data collection process as well as the strategies required to ensure that good quality data is collected need careful consideration. First is the need to ensure diversity of corpus content even within the spoken genre, that is, to ensure that the corpus covers different registers, styles, and genres. In a bid to address the problem of coverage, corpus developers may resort to intentionally asking and probing speakers to produce particular kinds of data that may be lacking.

There is also the need to ensure the quality of recordings. The audio-recording activity should be taken seriously, and not as a less significant task only important in facilitating language transcription and analysis as some researchers would like to think. It is necessary to train fieldworkers who should approach this task with the required skill and thoughtfulness. Also critical is the need to use high-quality recording devices since the success of the transcription task largely depends on the quality of recordings.

Corpus developers should also seriously consider the issues of protocol and research ethics. This is mainly because spoken data directly captures and represents individual study participants in ways that are different from written data. The data, especially from conversations, should be fully anonymised and study participants should be guaranteed this anonymity. For understudied languages whose communities are often under a wide range of socio-economic and political pressures, means and ways of dealing with sensitivities should be devised, both in accessing and distributing orally collected data.

In instances where published materials are available, corpus designers should consider forming partnerships with publishers, the copyright holders, to deal with issues of access and distribution of corpus materials. Arrangements can also be made to access data from translators and radio programmers through agreements accompanied by due acknowledgment. For printed texts that may be available, scanning the texts using an optical character recognition (OCR) programme may be more viable compared to typing.

4.3 Transcription. After recording, audio texts should be transcribed. The transcription choice is often between orthographic and phonetic methods and depends on the corpus designers' preference. Gerstenberger, Partanen, Rießler, and Wilbur (2017) contend that orthography-based transcription allows for quicker and more efficient transcriptions. They argue that this transcription method also provides for easy incorporation of already available digitised or printed texts into the corpus. Another advantage of orthographic transcription is that it makes the data usable to a wider variety of potential users. That is, orthographic transcription makes corpus texts accessible to language communities since general users are used to reading in orthography. This also makes it easier to hire research assistants to help with the transcription of audio data. This is unlike phonetically transcribed data that can only be produced and accessed by those with some training in linguistics.

However, a lot needs to be seriously considered in relation to the use of orthographic conventions in transcribing spoken data in understudied languages. This is especially so given the fact that whilst most of these languages do not have standard orthographies, others have multiple or divergent writing systems. Divergence or inconsistency in orthographies is seen in numerous cases where different writers tend to write differently in most of these languages. Such cases call for the standardization of writing systems first before they can be used as the basis for transcription. Only a standardized spelling and punctuation system can produce a consistent and thus more readily and effectively searchable corpus (Cox, 2010:256). Tools that rely on this uniformity can also easily annotate standardized data. For those languages where written materials exist, orthographic standardization may entail a process of correcting already existing written materials so that they conform to the new or selected orthography for corpus construction.

4.4 Tagging. To increase the utility of corpora, be it in well-studied or understudied languages, the texts should be annotated. In fact, for a corpus to be usable for language technology and linguistic research, it should be accurately segmented, lemmatised, and tagged in ways that make it an efficient source of data for a variety of uses. Useful tags or identification codes should be put on the texts if data should be retrievable in its specificity. In this case, the corpus designer should anticipate the kinds of data that users may want to find in the corpus so that the tagging process is guided by the users' needs. The texts should be marked up with header items and other text elements as required by language engineers. Some of the text elements that can be marked in order to increase efficiency in data retrieval include paragraphs, sentences, headings, foreign text, and parts-of-speech. For spoken texts, corpus designers can also mark utterances or speaker turns, false starts, repetitions, self-corrections, incomprehensible speech, and foreign speech segments, amongst other markings. A wide range of sociolinguistic information such as age, gender, education and community position of the interviewees and other information such as place and date of interview can also be collected and marked.

Of importance to note is the fact that the richness of tagging is normally proportionate to the potential value of the materials to users. For example, a well-tagged corpus enables researchers such as lexicographers and grammarians to see word searches as well as summaries of words' grammatical and collocational behaviours. A well-annotated corpus can also be used to develop language-specific tools such as part-of-speech taggers, syntactic analysers, lemmatisers and morphological parsers in a more efficient manner. Text annotation should enable a variety of search options and where necessary, it should make it possible for researchers to create sub-corpora from the main corpus. Vinogradov (2016, 132) notes that such corpus characteristics make a corpus suitable for a wide range of research questions. Given the state in which the understudied languages of Zimbabwe are found, it would be beneficial if tag sets developed for their annotation could be designed in ways that make them usable for a wide range of applications.

4.5 Corpus size. The issue of corpus size is a contentious one. Whilst it is generally agreed that large corpora (both in terms of the total number of tokens in the corpus and the types of words in it) are desirable for adequate linguistic analysis, the minimum number of words for such corpora has been hard to pin down. Some scholars have been brave enough to consider a corpus of at least one million running words as adequate for analytical purposes (see, for example, Nathan, 2009:104). Kennedy (1998) also alludes to the fact that for first-generation corpora, one million words was considered huge. However, the question of size as a determining factor on the reliability of a corpus should not be a discouragement for those who can only manage to construct a 'small' corpus. This is because no corpus size can be considered big enough to adequately represent a language or its variety. Sinclair (1991:20) notes that even a corpus containing ten or twenty million running words might constitute 'a useful small general corpus' but 'will not be adequate for a reliable description of the language as a whole'.

Huge corpora running into many millions may be difficult to build in Zimbabwe's understudied languages. This is mainly because of the situation in which these languages are found that inhibit the amount of data that can be collected. Some of the inhibiting factors have already been identified as the lack of written materials, the small number of speakers in some languages, the moribund state of some of them, limited funding, and physical remoteness. Experience from the construction of corpora for Shona, Ndebele and Nambya through the ALLEX – ALRI research programme shows that even the one-million-word mark corpora considered adequate for first-generation corpora may be difficult to reach. Whilst non-availability of written texts in both printed and electronic form has been cited as the main drawback, this is also coupled with the high cost associated with the labour-intensive and time-consuming corpus development tasks such as audio data collection, transcription, annotation, and cleaning that have to be done manually in these languages where automation is not possible. Thus, whilst a billion-word corpus can easily be achieved in major languages through harvesting text on the internet and other data-rich sources, the same may not be said of the Zimbabwean languages.

From the foregoing, a proposal can be put forward to consider the construction of 'small' general language corpora of the monitor type as a starting point. The corpora should be designed to characterize the state of the contemporary languages concerned in their various social and general uses. The starting point should not be issues to do with the availability or non-availability of texts, but the notion of a feasible language corpus.

4.6 Training of field and office workers. It is true that the outcome of corpus development is partly dependent on how the corpus developers deploy human knowledge, skills, and technologies. However, it has been observed that there are some tasks that linguists often take for granted. One such task that is taken light-heartedly is audio recording, which Nathan (2009) bemoans as usually poor in the collection of primary data for linguistic analysis. Perhaps that is the reason why Dietrich Schüller has described the audio recording methodology by linguists as one of the least scientific practices of all disciplines (Nathan, 2009:109). There is a need to properly train data collectors (especially in microphone handling, acoustics, and managing noisy recording environments, which are the greatest determinants of audio recording quality) and processors (for converting incoming materials into corpus format).

Following corpus development experience from the ALLEX-ALRI research programme, advanced linguistics undergraduate and postgraduate students who are native speakers of the respective languages can be recruited and trained to effectively undertake tasks such as audio data recording, transcription, annotation, and corpus normalization or editing. Similar training can also be extended to competent members of the community who are interested in the development of their languages.

5. Conclusion

The article highlighted the importance of corpora as sources of linguistic evidence in the study of language in general and in the study of understudied languages in particular. It also noted the significance of corpora in the development of language engineering tools that assist researchers in corpus querying. The article also examined issues of corpus designing at both the theoretical level as well as its practical application in understudied languages. Taking Zimbabwe's marginalised languages as a case in point, an observation was made that although it is desirable to build huge and representative corpora in all languages, there are quite a number of challenges that militate against this ideal in understudied and under-documented languages that generally lack sufficient materials required to construct such corpora. A proposal was made to construct 'small', general reference, and expandable corpora as a starting point. The corpora should be linguistically marked up so that they are useful in the development of reference works such as normative grammars and dictionaries. The same corpora would in turn be valuable in the development of language-specific engineering tools that make it easier and faster to search, sort, and manipulate corpus data.

References

Biber Douglas, Susan Conrad and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge: University Press.

Cox Christopher. 2010. Probabilistic tagging of minority language data: A case study using Qtag. In: Stefan Th. Gries, Stefanie Wulff and Mark Davies (eds.). *Corpus linguistic applications: Current studies, new directions*. Amsterdam: Rodopi. pp 213-231.

De-Schryver Gilles-Maurice and Danie J. Prinsloo. 2000. The compilation of electronic corpora, with reference to African languages. *Southern African linguistics and applied language studies 18*. pp 89-106.

Gerstenberger Ciprian, Niko Partanen, Michael Rießler and Joshua Wilbur. 2017. Instant annotations – applying NLP methods to the annotation of spoken language documentation corpora. *The 3rd international workshop for computational linguistics of Uralic languages.* St. Petersburg: Association for Computational Linguistics. pp 25–36.

Himmelmann Nikolaus. 2006. Language documentation: What is it and what is it good for? In: Jost Gippert, Nikolaus Himmelmann and Ulrike Mosel (eds.). *Essentials of language documentation.* Berlin: Mouton de Gruyter. pp 1-30.

Leech Geoffrey. 1992. Corpora and theories of linguistic performance. In: Jan Svartvik (ed.). *Directions in corpus linguistics: Proceedings of nobel symposium 82, Stockholm, 4-8 August 1991*. Berlin: Mouton de Gruyter. pp 105-125.

Kennedy Graeme. 1998. *An introduction to corpus linguistics.* London and New York: Longman.

Maxwell Mike and Baden Hughes. 2006. Frontiers in linguistic annotation for lower-density languages. *Proceedings of the workshop on frontiers in linguistically annotated corpora.* (July 2006, Sydney, Australia): pp 29-37.

McEnery Tony, Paul Baker and Lou Burnard. 2000. Corpus resources and minority language engineering. In: M. Gavrilidou, G. Carayannis, S. Markantontou, S. Piperidis and G. Stainhauoer (eds.). *Proceedings of the second international conference on language resources and evaluation.* LREC: Citeseer. pp 1-6.

McEnery Tony and Nick Ostler. 2000. A new agenda for corpus linguistics – working with all of the world's languages. *Literary and linguistic computing*, 15: 403-418.

McEnery Tony and Andrew Wilson. 2001. *Corpus linguistics: An introduction. 2nd* Edition. Edinburgh: Edinburgh University Press.

Midrigan-Ciochina Ludmila, Victoria Boyd, Lucila Sanchez Ortega, Diana MalanceaMalac, Doina Midrigan, David P. Corina. 2020. A representative corpus of the Romanian language: Resources in underrepresented languages. *Proceedings of the 12th conference on language resources and evaluation (LREC 2020)*, pp 3291–3296.

Nathan David. 2009. Audio responsibilities in endangered languages documentation and archiving. In P.K. Austin (ed.) *Language documentation and description, Vol 6*. London: SOAS. pp 101-116.

Sinclair John. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sinclair John. 2005. Corpus and text – basic principles. In: M. Wynne (ed.). *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books. pp 1-16.

Stubbs Michael. 2001. *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Szymanski Terrence D. 2011. Morphological inference from Bitext for resource-poor languages. Doctoral Dissertation. University of Michigan.

Vinogradov Igor. 2016. Linguistic corpora of understudied languages: Do they make sense? Káñina 40 (1): 127-141.

Walther Géraldine and Benoît Sagot. 2017. Speeding up corpus development for linguistic research: Language documentation and acquisition in Romansh Tuatschin. *Proceedings of the joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature.* Vancouver, BC: Association for Computational Linguistics. pp 89–94.

Emmanuel Chabata
Department of Languages Literature and Culture
University of Zimbabwe
emmanuelchabata@yahoo.com