Spoken word corpus and dictionary definition for an African language

Wanjiku Nganga, Ikechukwu Achebe

Journal of Data Mining and Digital Humanities

Spoken word corpus and dictionary definition for an African language

Wanjiku Nganga[1], Ikechukwu Achebe[2]

1 University of Nairobi, Kenya

2 Igbo Archival Dictionary Project, Nigeria

Abstract
The preservation of languages is critical to maintaining and strengthening the cultures and identities of communities, and this is especially true for under-resourced languages with a predominantly oral culture. Most African languages have a relatively short literary past, and as such the task of dictionary making cannot rely on textual corpora as has been the standard practice in lexicography. This paper emphasizes the **significance of the spoken word and the oral tradition as repositories of vocabulary, and argues that spoken word corpora greatly outweigh the value of printed texts for lexicography. We describe a methodology for creating a digital dialectal dictionary for the Igbo language from such a spoken word corpus. We also highlight the language technology tools and resources that have been created to support the transcription of thousands of hours of Igbo speech and the subsequent compilation of these transcriptions into** an **XML-encoded textual corpus of Igbo dialects. The methodology described in this paper can serve as a blueprint that can be adopted for other under-resourced languages that have predominantly oral cultures**.

keywords
audio corpus; dictionary definition; digital resources; Igbo; oral tradition; under-resourced languages

INTRODUCTION

The preservation of languages is critical to maintaining and strengthening the cultures and identities of communities. Recorded and printed documentation, preserved via durable physical and digital media, is the typical way in which the sounds and usage of languages are preserved. Dictionaries, print or digital, play an important role in the preservation, revitalization and maintenance of languages, and this is particularly pertinent for communities with largely oral traditions because such cultural heritage is threatened with extinction following the inevitable death of older speakers. The task of lexicography is complicated for predominantly spoken rather than written languages, due to the absence of a large body of print-texts created over a substantially long period of time from which items and their contextual meanings can be derived or deduced. Lexicography is virtually impossible without a linguistic corpus. For many of the world's languages, for example, printed texts have been the basis for such a corpus. From such corpora, it was possible to make a census of a language's discrete wordforms. In addition, the history of the meanings and morphological status of such forms can be easily derived from those period-specific textual records. Scholars for most African languages like Igbo that do not have a long history of written texts, would

have to base their analysis of language-use on an actual and verifiable body of spoken evidence established and recorded in advance of specific citations.[1]

The work presented in this paper documents a tested methodology for defining dictionaries for languages that are predominantly spoken rather than written, and where the vast proportion of the language remains undocumented. This approach, undertaken for the Igbo language, proceeds from a fundamental lexicographical principle: that speech occupies a much greater role in language-use than writing; and that for many African languages, with relatively short literary histories, the significance of the spoken word and the oral tradition as repositories of vocabulary, greatly outweighs the value of printed texts for lexicography.

I. A BRIEF HISTORY OF DICTIONARY MAKING IN IGBO

The Igbo language is one of Africa's great indigenous languages, with over 30 million speakers around the world. Igbo belongs to the Benue-Congo branch of the Niger-Congo language family of Africa and is spoken in seven states in Nigeria—Abia, Anambra, Delta, Ebonyi, Enugu, Imo and Rivers, as well as among large and growing émigré populations in the United States and Europe. Linguists recognize more than fifteen Igbo dialects in existence[2] but studies in Igbo dialectology are ongoing and the final number is likely to be higher.

The history of dictionary-making in Igbo may be said to have begun in the 18th century with the production of bilingual wordlists and glossaries by European missionaries. In 1777 for instance, the Moravian mission agent G. C. A. Oldendorp published *Geschichte der Mission der evangelischen Brüder…*, which contained a number of Igbo words and numerals. The production of bilingual wordlists and vocabularies by European explorers and missionaries in Africa continued in the first half of the 19th century; and they were published either in separate volumes or included as appendices or glossaries at the end of grammar books. Samuel Crowther, a native agent of the Church Missionary Society (CMS), produced the Isoama-Ibo primer in 1857; and a number of translations of the Gospel into Igbo followed Crowther's primer during the second half of the 19th century. The documentation of Igbo vocabulary at this time owed a great deal to the pragmatic concerns of the compilers, who were invariably Christian missionaries actuated by evangelical zeal; and their works lacked semantic coverage and basic phonological, morphological and syntactic information. These pioneers were more concerned with translating Western religious texts into Igbo, than in documenting Igbo as it was spoken.

With the 20th century came the publication of dictionaries [Ganot, 1904], as well as a number of scriptural texts written in what was then called 'Union Ibo'—an expedient and ultimately unworkable hybrid—invented by the CMS under the British colonial administration for the evangelical mission. In 1913 for instance, the Union Ibo Bible was published, produced by Archdeacon T. J. Dennis. Since then, several other dictionaries and word lists have been compiled by [Ogbalu, 1962], [Williamson, 1972], [Anoka, 1979], [Nnaji, 1985], [Echeruo,

[1] In 2001, M. J. C. Echeruo, *William Safire Professor of Modern Letters* at Syracuse University, coordinated the initial workshops at which issues of the project's macro-structure were first raised and agreed upon. The views expressed in this paragraph are the results of those deliberations.

[2] The dialects of Igbo include Afikpo, Aniocha, Azumini, Bonny/Opobo, Echie, Egbema, Igbouzo, Ngwa, Nsa, Mbaise, Nnewi, Nsuka, Oguta, Oka (Awka), Ohuhu, Onitsha, Orlu, Owerri, Umuahia, Unwana and Uturu.

1998] and [Igwe, 1999]. These wordlists and dictionaries were produced single-handedly, by individual scholars or enthusiasts, without collaboration among lexicographers, linguists and others. Consequently, one of the major inadequacies of these dictionaries for humanities studies has been that, like all previous efforts at documenting the language, these dictionaries are limited in the scope and coverage of the vocabulary of Igbo language.

II. DEFINING A DIGITAL DIALECTAL DICTIONARY FROM A CORPUS OF SPOKEN IGBO

2.1 Creating a pan-Igbo speech corpus

To address the limitations of existing Igbo dictionaries and to bring the development of Igbo lexicography in line with international standards, it was necessary to devise an approach that would espouse lexicographical principles and standards, but would also recognize the foundational role of a spoken corpus for dictionary definition, by working from the oral tradition of languages such as Igbo. The first step therefore in creating a dialectal dictionary of Igbo was to record spoken Igbo in all its dialects, establishing an accurate, verifiable corpus of spoken Igbo that would form the basis for dictionary definition. This enormous task was undertaken between 2000-2020 by the Igbo Archival Dictionary Project (IADP), an association of linguists, anthropologists, historians, computer scientists and other scholars, established with the purpose of conducting research and salvage work to preserve and develop the Igbo language. The task of creating a comprehensive dictionary of the Igbo language is a large one, because it entails an inquiry not simply into a single dialect, but into the entire complex of Igbo dialects. As the first major effort in documenting the Igbo language on a large scale, the IADP, working at seven Nigerian universities, has trained and deployed more than fifty fieldworkers, linguists and consultants to work in Nigeria on the project over a period of 20 years. This effort has yielded more than 1000 hours of local language speech from hundreds of interviews conducted across Igbo towns and villages, making it one of the largest salvage projects ever undertaken for documenting an African language from speech.

2.1.1 Creating a pan-Igbo text corpus through transcription

Working from this pan-Igbo audio corpus, the next step was to transform this corpus into its text equivalent to support the lexicographic work of defining a dialectal dictionary. This was done by creating digital transcriptions of the audio recordings, and ordering these into an XML-encoded text corpus. The first challenge that needed to be addressed for transcription was that of rendering Igbo orthography digitally, since Igbo contains many diacritical marks for tone-marking of vowels and nasals (high, low and downstep) as well as special characters that are not directly accessible from a standard keyboard. Further, the research into Igbo dialects had identified sounds that previously had not been documented and written, and these were added into the composite synchronic alphabet of Igbo dialects by [Achebe et al, 2011] and then produced digitally as well. To facilitate easy and accurate typing of fully tone-marked Igbo texts, we developed Igbo software-based keyboards that enable a transcriber to insert all Igbo characters (both lower and upper case) and associated phonetic equivalents with requisite tonemarking diacritics, with a single keystroke. In addition, we created the Igbo Corpus Builder (ICB) software which is a special editor for creating interview transcriptions encoded in XML. The ICB facilitates the inclusion of interview metadata such as the interview date, location (state, local government area, town, village), interviewee details (age, gender, occupation), dialect information and discourse/topic classification. Text transcriptions

generated via the ICB are what constitute the pan-Igbo text corpus that is used for the lexicography task.

2.2 Lexicography based on pan-Igbo text corpus

The lexicography work proceeded by first pre-processing the interview texts to generate alphabetically sorted wordlists with associated concordance information for a given word, which provides the context for meaning identification and verification. We developed a suite of text pre-processing software tools that performs a myriad of natural language processing tasks including wordlist and concordance generation. Working with headwords identified from the pre-processed corpus texts, lexicographers are able to define dictionary entries in keeping with best practice.

Lexicographical work on a multi-dialectal language such as Igbo presents challenges at the macro-structure level; and dictionary compilation needs to address the question of how to represent headwords; that is, whether dialect forms should be selected as headwords and how variant forms should be represented. For example, if we take the Igbo variants in which are glossed as '*body*' in English, the lexicographer must decide whether to list these as variants or headwords, or to use only one dialect form.

Figure 1. Dialect variants for *body*

.

In addressing this problem, we take the view that a dialect dictionary basically works with three core data types: form, sense (meaning) and location, since it aims to document and classify dialectal forms or variants that are used in specific senses for specific locations. For many dialect dictionary projects, the choice of how to organize these core data types has largely been influenced by the publication medium: printed media are one-dimensional and linear, and therefore dialect dictionaries invariably present the data sequentially, according to some ordering principle. In practice, this means that the editors have to choose one of the core types as the most important organizing principle. However, it is clear that the choice of opting for one organization principle over the other is not based on fundamental differences between one core data type over the others, but purely on practical reasons. In other words, different uses of the data are better catered for by one or the other organizational principle; but the nature of the data does not have an intrinsic hierarchy such as "sense over form" or "form over sense". We adopted a dictionary organization model that allows us to abandon the distinction between macro- and microstructure, opting instead to reduce micro-structure to the relation between the three core data types, and to broaden macro-structure to a dynamic, use-driven classification that is based on a combination of the basic tripartite units.

To achieve this flexible data model, we ensured that in the design of the database schema the relationships inherent in the data were separated from the core data itself. This flexibility in working with the data makes it possible to organize the data in several different ways, allowing the user to choose the viewpoint most suitable to their needs. For example, if the user wants to know the form variation for the sense "body", s/he would choose a sense-based view, which would show all six variants given in above. If s/he wants to know what the sense distribution of the form "àkwá" is, s/he would choose the form-based view of the data. And finally, if s/he wants to view dictionary data for a particular dialect (location), for example to make a local dictionary, s/he would want to have a location based view of the data. By adopting a database design that allows for multiple views, the resulting dictionary can be used for many different purposes.

III. USING THE DIGITAL, DIALECTAL DICTIONARY OF IGBO

A digital repository offers a diverse array of accessibility options. To increase access to the Igbo dialectal dictionary, the final release will be hosted online, in the same way as is the prototype dictionary platform which is currently under development. Given the multilingual nature of a dialectal dictionary, it is important to ensure that the online platform enables users to search for words with flexibility, since different users have different needs and knowledge backgrounds. To this end, the digital dialectal dictionary has been designed to include features which enable a user to find the meaning(s) of a given word, but also to obtain further information on any dialectal variations that may be associated with the current search term. By considering the variability of written Igbo texts with respect to tone-marking, we anticipated different orthographical representations for search terms, where users can type words with or without tone-marking. The latter scenario could be occasioned by several factors such as lack of an appropriate keyboard that allows efficient tone-marking of Igbo words, lack of knowledge on the exact tone-pattern associated with the search word of interest, or an intentionally-unspecified (discovery) search.

Since Igbo words only bear meaning when tone-marked, we provide two ways to enable a user accomplish the task of looking up meanings successfully: i) an Igbo character palette that contains all Igbo characters with their tone-marked varieties and ii) automatic tone-marking, accomplished by a software module which automatically generates all tone-marked possibilities for a given canonical (i.e. non tone-marked) search term. The results of auto-tonemarking are filtered to reflect only those variations that have been defined in the dictionary database. These variations are provided as a list, from which the user can select the intended tone-marked search term. For each search term, a wide range of information is provided:

- English gloss, with a detailed English description where necessary.
- The search word with appropriate tone-marking—this is important for cases where the user searches for a canonical word, and are able to see how it is tone-marked for different meanings.
- Word syllabification and associated syllabic tones, since the syllable is the tone-bearing unit in Igbo.
- The phonetic spelling of the search word, using the International Phonetic Alphabet.
- A list of dialects where a given tone-marked variant bears the same meaning.
- A word's geographical distribution, which is availed interactively via a geographical information system (GIS) of Igbo land, making it possible for the user to see the exact

villages, towns and states where the word is spoken, and the variances that exist across meanings and dialects

- An audio pronounciation of the search word which is generated by a word-level Igbo speech synthesis engine, a feature that greatly enhances a user's experience, since the user can hear the tonal richness of Igbo dialects.

IV. CONCLUSION

This paper emphasizes the importance of creating dictionaries as part of language and cultural preservation and maintenance. It has also highlights the complexities of dictionary definition for languages with numerous, spoken dialects. We have described a methodology for creating a digital dialectal dictionary from an audio corpus. Our methodology highlights the challenges facing lexicographical work for languages that have a limited or non-existent literary tradition and which are characterized by limited textual and electronic resources. By recognizing that most African languages have a predominantly rich oral, but largely non-literary, past traditions, the work described here undertakes a massive task of creating an audio databank for one of Africa's biggest languages. By devising language technology and software tools for language processing, the paper describes the implementation of a workable, efficient and cyclic workflow that proceeds from audio recording of Igbo dialects, to subsequent digital transcription incorporating corpus encoding standards, in order to yield the first digital, dialectal, audio and text corpus of Igbo.

This corpus enables Igbo lexicographical work to proceed in keeping with international standards. One of the major achievements of the work undertaken so far has been in solving the orthographic challenges that have plagued the creation of Igbo corpora—that of creating a corpus of fully tone-marked texts. This has been achieved by way of developing Igbo software keyboards and the unicode-based ICB editor. This is a significant accomplishment, since non tone-marked corpora are useful perhaps only to native Igbo speakers who can decipher the intended meaning. In addition, the lack of tone-marking results in an explosion of ambiguity at different levels of grammatical analysis which greatly complicates any language technology efforts. We have also successfully developed language technology tools and a natural language processing pipeline that support the compilation of a corpus of fully tone-marked texts, processing these texts for different linguistic analyses, and displaying fully tone-marked text via a web-browser. This brings the Igbo language into the internet domain without any representation limitations. With these tools, the task of creating a comprehensive dictionary for Igbo dialects is now feasible.

4.1 Significance of the work for under-resourced languages

The work described here, though focussed on Igbo, serves as a blueprint that can be adopted for languages with similar contexts. In particular, we have shown how to document and preserve the cultural diversity of a language with numerous dialects, by defining a dialectal dictionary from a spoken corpus. Further, we have demonstrated that such strides can only be attained when multidisciplinary expertise in linguistics, culture, anthropology and language technology are leveraged to propose and implement novel approaches and solutions that work for under-resourced languages from an oral, largely non-literary, tradition.

References

Achebe I., Ikekeonwu C., Eme C., Emenanjo N., and Ng'ang'a W. *A Composite Synchronic Alphabet of Igbo Dialects (CSAID)*. IADP (Awka, 2010; New York, 2011).

Anoka, K. et al. *Pronouncing Dictionary of Igbo Place Names*. Owere: Culture Division, Ministry of Information, Culture/Youth and Sports, 1979.

Echeruo M.J.C. *Igbo-English Dictionary, with an English-Igbo Index*. Yale University Press (New Haven-London), 1998.

Ganot A. *English-Igbo-French Dictionary. Sodality of St. Peter Claver, Onitsha Dialect*. Onitsha/Rome, 1904.

Igwe G.E. *Igbo-English Dictionary*. University Press Plc (Ibadan), 1999.

Landau S.I. *Dictionaries: The Art and Craft of Lexicography*. Cambridge University Press (Cambridge, UK), 2001.

Nnaji, H.I. *Modern English-Igbo Dictionary*. Onitsha: GONAJ Books, 1985.

Ogbalu, F. C. *Okowa-Okwu: Igbo-English-English-Igbo Dictionary*. Onitsha: Varsity Printing Press, 1962.

Williamson K. *Igbo-English Dictionary*. Ethiope Publishing (Benin City, Nigeria), 1972.