Towards a Comprehensive, Machine-readable Dialectal Dictionary of Igbo

Wanjiku Ng'ang'a

School of Computing and Informatics
University of Nairobi
Kenya
wanjiku.nganga@uonbi.ac.ke

Abstract

Availability of electronic resources, textual or otherwise, is a first step towards language technology research and development. This paper describes the acquisition and processing of a multi-dialectal speech and text corpora for the Igbo language. The compiled corpus provides the key resource for the definition of a machine-readable dialectal dictionary for Igbo. The work centres around an online portal that facilitates collaborative acquisition, definition and editing of the dialectal dictionary. The complete dictionary, which includes features such as phonetic pronounciation, syllabification, synthesized pronounciation as well as GIS locations, is then made available via a website, which will be accessible publicly once sufficient entries have been defined.

1. Background

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 5 major dialect "clusters" and more than 15 main Igbo dialects in existence, but studies in Igbo dialectology are ongoing and the final number is likely to be higher (Achebe and Ng'ang'a, 2007).

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 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 (Crowther, 1882); and a number of translations of the Gospel into Igbo followed Crowther's primer during the second half of the 19th century (Norris, 1841; Schon, 1861; Koelle, 1854). 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 (Emenanjo and Oweleke, 2007). 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 which was 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 (Williamson, 1972; Echeruo, 1998; 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 (Emenanjo and Oweleke, 2007).

1.1. The Igbo Archival Dictionary Project

Against this backdrop, there was an urgent need for the collaboration of lexicographers, linguists, interdisciplinary scholars, language planners and other stakeholders to bring the development of Igbo lexicography in line with international standards. In 1999 the Nigerian writer, Chinua Achebe, delivered a lecture in which he called on linguists and other scholars to check the declining fortunes of the Igbo language by working towards its comprehensive documentation. In his speech, he noted that "*Language is not a piece of iron that the blacksmith takes and puts into the fire, takes out and knocks into shape or moulds as it pleases him. Language is sacred, it is mysterious, it is fearsome. It is living and breathing. It is what separates humans from animals. It separates people from their counterparts, towns from their neighbors*", emphasizing the need to develop a comprehensive dictionary of Igbo and all its dialects.

Following from this, Ike Achebe founded the Igbo Archival Dictionary Project (IADP) in 2001 as an association for leading linguists, anthropologists, historians and other scholars on Igbo language and culture 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 into a single dialect, but into the

entire complex of Igbo dialects. In 2002, IADP began the first significant effort in the history of the Igbo language, for its scientific documentation on a large scale. Over the subsequent years, IADP has been working at seven Nigerian Universities laying the ground work for a massive effort at Igbo documentation via the recording of spoken Igbo in all its dialects. IADP has trained and deployed more than 50 fieldworkers, linguists and consultants to work in Nigeria on the project. The project has to-date conducted fieldwork and recorded more than 1000 hours of local language speech in Igbo villages and towns, making it one of the largest salvage projects ever undertaken for documenting an African language from speech. Since 2002 the IADP has been building its Igbo Corpus from transcriptions of the audio recordings in its archives, with an objective of building the Igbo Corpus from local language speech to at least 25 million words by the end of 2012 (Achebe and Ng'ang'a, 2007).

1.2. The Igbo Online Resources Project

The Igbo Online Resources Project (IORP) is a sub-project within the greater IADP whose aim is to develop electronic resources for Igbo. The ultimate goal of the IORP is to develop language technology resources and tools that will facilitate the development of end-user applications for Igbo such as Word processors, Spell-checkers, Speech-mediated interactive voice response systems, and Machine translation of Igbo texts to name a few. To achieve this, we have embarked on the creation and development of varied electronic resources which include textual and audio corpora, a comprehensive machine-readable dictionary, speech synthesis software and software tools to perform miscellaneous linguistic tasks such as tone-marking and word syllabification. It is envisaged that as this repertoire of language technology tools and resources continues to grow, the IORP will be on course to achieve her overall objectives, as stated earlier. This paper describes the methodology and computational efforts employed by the IORP in delivering a web-based Igbo dictionary.

2. Creating the Dialectal Dictionary

IORP proceeds from a fundamental lexicographical principle: that speech occupies a much greater role in language-use than writing; and that for 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 (Achebe and Ng'ang'a, 2007). Citation files have traditionally been based on writing; but it is now clear to lexicographers that the traditional citation file cannot adequately represent the various uses of language (Landau, 2001), especially for a language such as Igbo that has a predominantly non-literary past. This is because the vast proportion of the Igbo language is still undocumented, and Igbo continues to be a severely under-resourced language. One of the problems associated with lexicography for many African languages is 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 European 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 word-forms. In addition, the history of the meanings and morphological status of such forms can be easily derived from such period-specific textual records. Scholars for languages like Igbo that do not have a long record of written texts, will have to base analysis of language-use on an actual and verifiable body of spoken evidence established and recorded in advance of specific citations[1] (Achebe and Ng'ang'a, 2007). The transcriptions of the audio recordings collected by the IADP form the key resource for dictionary definition for the IORP, as explained in subsequent sections. Delivering a web-accessible Igbo dictionary involves a 6-step process that manipulates/produces the following resources:

1. An Audio record of an interview is transcribed to produce
2. An XML-formatted transcript of the interview. This is then uploaded to a central server. Several such transcriptions form the
3. Electronic Igbo Corpus. Various pre-processing modules are applied to the corpus yielding
4. Unique word lists and concordances. These are used during the lexicographic work that produces
5. Unmoderated dictionary entries. Once these are moderated by the editorial committee, a list of
6. Moderated dictionary entries are then available for release via the online dictionary.

2.1. Creating the electronic Igbo Corpus

To avail the audio corpus of the IADP for lexicographic work, the first task is that of creating electronic transcriptions of the audio recordings. The first challenge that was encountered was that of orthography 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. 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) with requisite tone-marking diacritics, with a single keystroke. In addition, we have created the Igbo Corpus Builder (ICB) software

[1]Members of the Igbo Archival Dictionary Project continue to produce scholarship on aspects of the work they have done for the project. In 2003, Ozo-Mekuri Ndimele, IADP's coordinator for the Echie dialect in 2002, published "A Concise Grammar and Lexicon of Echie" (Aba: National Institute for Nigerian Languages). Similarly, E. N. Oweleke, a foundation member of the IADP, who has followed closely the development of Igbo lexicography in IADP submitted a PhD dissertation titled, "Some Issues in Lexicography and the Problems of Affixation and Verb-Noun Selectional Restrictions in the Igbo Dictionary": (Unpublished Ph.D. Dissertation, 2007), University of Port Harcourt, Nigeria.

which is a special editor for creating interview transcriptions encoded in XML. The ICB facilitates insertion of interview metadata which includes the interview date, location (State, Local government area, town, village), interviewee details (age, gender, occupation), dialect information and discourse/topic classification.

2.2. The Lexicographical Task

Once an interview transcription is created via the ICB, it is added to the electronic textual Igbo corpus. The next step is that of pre-processing the interview texts to facilitate lexicograhical work, for which we developed a suite of pre-processing software tools that perform a myriad of natural language processing tasks. Text pre-processing mainly involves obtaining an alphabetically sorted list of all unique words in the text. A concordance for each unique word is also created as this provides the context for meaning identification and verification. Using this information, the lexicographers are then able to define dictionary entries using the headwords identified from the texts. 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 Table 1 which are glossed as 'body' in English, the question is, should these be listed as variants or headwords or should the lexicographer decide to use one dialect form? (Emenanjo and Oweleke, 2007).

àshụ (in Enuanị)
àrụ (in Onịcha)
ẹhụ (in Ika)
ẹsụ (in Ụkwụani)
ẹshụ (in Nsụka)
àhụ (in Owere)

Table 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 form variants that are used to talk about specific senses in 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 over the other is not based on fundamental differences in importance of one core data type over the others, but purely on practical reasons. Different uses of the data are better catered for by one or the other organization, but the nature of the data does not have an intrinsic hierarchy such as "sense over form" or "form over sense".

We have adopted a dictionary organization model that allows us to abandon the distinction between macro- and micro-structure, 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 (Achebe and Ng'ang'a, 2007). To achieve this flexible data model, we ensured that the relationships inherent in the data were separated from the core data itself, in the design of the database schema. This flexibility in working with the data, makes it possible to organize the data in several different ways, thereby 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 will choose a sense-based view which would show all six variants given in Table 1. If s/he wants to know what the sense distribution of the form "àkwá" is, s/he will want to have the form-based view on 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 will want to have a location-based view on the data. By adopting a database design that allows for multiple views, the resulting dictionary can be used for many different purposes.

The Igbo lexicographical task is a highly collaborative one, engaging different teams with different specialities: field workers, stenographers, transcribers, lexicographers and the editorial board. While most of the teams are within physical reach in Nigeria, the team of lexicographers and the editorial board are variously located across the globe. The IORP therefore developed an online portal to facilitate communication, data sharing and collaboration. The portal provides web-accessible tools to send and receive data files, browse the electronic corpus and generate corpus statistics, view the word lists and associated concordances, define words for the dictionary and moderate entries for release in the online dictionary. The portal creates a seamless collaboration platform and includes a wiki site to enhance communication and discussion amongst all the researchers.

2.3. Online Igbo Dictionary

After a lexicographer has completed defining a headword, the word must go through a moderation process, which is done by an editorial board, via the portal. The moderation ensures the accuracy of the entire entry. Once a word has passed the moderation stage, it becomes available for release in the online dictionary. Given that the overall objective is to develop a comprehensive, dialectal dictionary, the online dictionary provides features which enable a user to not only find the meaning(s) of a given word, but 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 anticipate different orthographical representations for search terms, where users may type words with or without tone-marking. The latter scenario may be occasioned by several factors such as lack of an appro-

priate keyboard that allows efficient tone-marking of Igbo words, lack of knowledge on the exact tone-pattern associated with the search word, 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[2] 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 (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. The user is however at liberty to search for any word, tone-marked or canonical, to find out if it is exists in the dictionary. Where a word does not exist, the website includes a "Suggest a Word" interface where users can suggest words for inclusion into the dictionary. Suggested words will go through an editorial and moderation process before being included into the dictionary.
For each search term, a wide range of information is provided:

1. English gloss, with a detailed English description where necessary.
2. The search word with appropriate tone-marking. This is important for cases where the user searches for a canonical word, and they are then able to see how it is tone-marked for different meanings.
3. Word syllabification and associated syllabic tones, since the syllable is the tone-bearing unit in Igbo.
4. The phonetic spelling of the search word, using the International Phonetic Alphabet.
5. A list of dialects where a given tone-marked variant bears the same meaning.
6. A word's geographical distribution which is availed interactively via a geographical information system (GIS) of Igbo land. Here, the user can see the exact villages, towns and states where the word is spoken, and with what variances across meanings and dialects.
7. An audio pronounciation of the search word which is generated by an Igbo speech synthesis engine which has been developed by the IORP[3]. This feature greatly enhances a user's experience as they can hear the tonal richness of Igbo.

2.4. Summary Statistics and Availability

To date, the IADP has collected over 1000 hours of Igbo speech. Of these, approximately 100 hours have been manually transcribed. Electronic transcription using the ICB is currently on-going. The project portal is complete with all dictionary definition modules having been tested and commissioned for use. The IORP is now in the electronic corpus aggregation phase where the ICB formatted, XML encoded interviews are being uploaded onto a common server-based data repository. Once a sizeable corpus has been obtained, the lexicographical work will begin in earnest. Currently, only a few entries have been defined, moderated and availed via the online dictionary.

All the digital resources developed by the IORP will be made available online - the online Igbo-English dictionary, the speech synthesis system, our GIS-based dialect maps and the machine translation engine will be launched as a suite of applications from a single web portal. IORP's preference for Open Source ware in the development of the products will ensure easy access, usability and wide dissemination via the Internet.

3. Conclusion

This paper has described a methodology for creating a machine-readable dictionary from an audio corpus. It highlights the challenges facing lexicographical work for African languages which are characterized by limited electronic resources, if any. By recognizing that African languages are largely oral with little or no literary past, 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 IORP implements a workable, efficient and cyclic workflow that transforms the audio corpus into an electronic textual corpus, adopting corpus encoding standards. With the corpus and requisite concordance tools in place, the lexicograhical task for Igbo is greatly enhanced if not simplified. In addition, by collecting spoken Igbo from across Igbo land, we are in a position to define the first, truly comprehensive dialectal dictionary of Igbo.

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 the software keyboards and the unicode-based ICB editor. This is a significant accomplishment since non tone-marked corpora are only useful to native Igbo speakers who can decipher the intended meaning (Uchechukwu, 2004a; Uchechukwu, 2004b). 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, bringing the Igbo language into the internet domain without any representation limitations. With these tools, the task of creating a comprehensive dictionary for Igbo is now a feasible reality.

[2]This palette is accessible on the dictionary page as a selectable component.

[3]IORP commissioned Dr. Chinyere Ohiri-Aniche to document Igbo phonetic units in 2007 as part of IORP's endeavour to build linguistic resources for Igbo.

4. Acknowledgements

This paper has benefitted from ongoing collaborative research and discussion within the Igbo Archival Dictionary Project. The author wishes to acknowledge particularly the contributions of Professor E. Nolue Emenanjo, the Chief Editor of IADP; Professor Clara Ikekeonwu, the Chairman of the Editorial Board of IADP; and participants at the recent IADP Coordinator's Workshop held in Awka on March 4-7, 2010, for insights on Igbo dialectology. Late Professor Adiele Afigbo, National Coordinator of IADP from 2006-2009, helped develop the methodological approach to fieldwork discussed in this paper. Professor M. J. C. Echeruo coordinated the initial workshops at which issues of IADP's macro-structure were first raised and agreed upon. We are grateful for discussions with Professor M.C. Onukawa, Professor Ozo Mekuri Ndimele, Professor Ezikeojiaku, Dr. E. Oweleke, Dr. Ceclia Eme, Dr. Chris Agbedo, Dr. B. Mmadike, Jumbo Ugoji and George Iloene, which have enriched this paper.

The work accomplished by the IADP and IORP has been supported by the World Bank, the United Nations Foundation, The Ford Foundation and the Fulbright Program. We are grateful to the Vice-Chancellors of Nnamdi Azikiwe University Awka, Imo State University Owerri, Abia State University Uturu, Ebonyi State University Abakiliki, University of Port Harcourt, and University of Nigeria Nsukka, who have actively supported the goals of the IADP. We also acknowledge the technical support of teams situated at Bard College, USA and University of Nairobi, Kenya.

5. References

Achebe, I. & Ng'ang'a, W. (2007). The making of the Igbo online resources project. Technical report, IADP, New York.

Crowther, S.A. (1882). *Vocabulary of Ibo Language*. London, UK: Society for Promoting Christian Knowledge.

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

Emenanjo, E.N. & Oweleke, E.N. (2007). Compilation of the Igbo archival dictionary. Technical report, IADP, Port Harcourt.

Ganot, A. (1904). *English-Igbo-French Dictionary*. Rome, Italy: Sodality of St. Peter Claver.

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

Koelle, S.W. (1854). *Polyglotta Africana or a Comparative Vocabulary of Nearly Three Hundred Words and Phrases in More Than One Hundred Distinct African Languages*. London UK: Church Missionary House.

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

Norris, E. (1841). *Outline of a Vocabulary of a Few of the Principal Languages of Western and Central Africa, Compiled for the Use of the Niger Expedition*. London, UK.

Schon, J.F. (1861). *Grammatical Elements of the Ibo Language*. London, UK.

Uchechukwu, C. (2004a). The Igbo corpus model. In *Proceedings of the 11th EURALEX International Congress*. Lorient, France: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

Uchechukwu, C. (2004b). The representation of Igbo with the appropriate keyboard. In *International Workshop on Igbo Meta-Language*. Nsukka, Nigeria: University of Nigeria.

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