

# The Igbo Corpus Model

Chinedu Uchechukwu\*

The Igbo language (also formerly written as Ibo) is spoken by about 20 to 25 million people (Eze and Manfredi 2001/2; Igboanusi and Peter 2005: 59) in the following states of South-Eastern Nigeria: Abja, Anambara, Ebonyi, Enugu, Imo States; and also in some local governments of Rivers and Delta states. It is one of Nigeria's three major languages and is used in the educational sector. It can also be studied at a degree level at different Nigerian universities. The language belongs to the Niger-Congo language group, where it is further classified as belonging to the Benue-Congo family (Bendor-Samuel 1989), or to the West Benue-Congo family (Willisamson & Blench 2000). The *Igbo Corpus Model* presented here is an ongoing project whose initial aim was to develop a corpus model of written Igbo texts. The work so far has been able to work out some basic issues on text typology, POS-Tags, and also identify some serious problems for Igbo corpus development. I shall summarize these points below.

## 1. Text Typology

This involved three different stages: establishing the available titles, categorizing them as literary genres, and categorizing them in an electronic format.

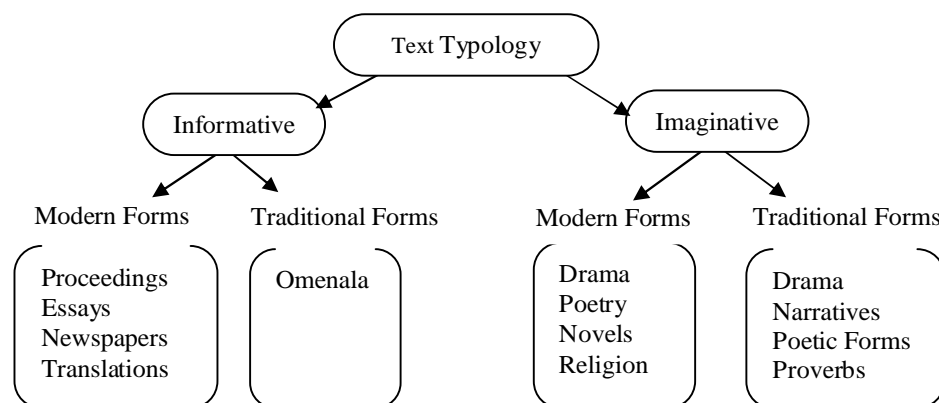
The first stage involved working with two Igbo scholars, Prof. Emenanjo and Dr. Osuagwu, to establish the available titles of written Igbo text up to the year 2004. This was not really easy, as some titles were either published only once, were self published, or simply available as typewritten manuscripts. Right from the early missionary activities to the present, Igbo studies has always been corpus based. This is as a result of the many dialects of the language which made it mandatory to compare data between dialects. However, the early works had only 'paper' as its storage medium. The disadvantage is that many of the gathered materials are sometimes lost or bleached due to the disintegration of the storage medium, as a result of which later researchers have to start

---

\* Otto-Friedrich-Universität, Bamberg  
neduchi@netscape.net

their own work from scratch. The *Igbo Corpus Model* is to contribute to putting the data in such a format that it would be permanently available to researchers.

The second stage involved categorizing the texts as literary works. This became necessary, as some Igbo texts were written on the basis of the traditional oral narratives, while others were written in line with the modern Western literary genres. Finally, there were also texts arising from the print media. Based on these facts, I divided the Igbo texts into two major groups: (1) the Informative and (2) the Imaginative. While this grouping should not be seen as an absolute or rigid classification, it helps to give an overview of the available literature. Each of the groups are further divided into (1) Modern Form and (2) Traditional Form. The Modern Forms refer to the different forms of writing that have some Western genres as their model, while the Traditional Forms are made up of the different forms of oral information or genres that had always existed in Igbo tradition and culture. The picture below gives an overview of the typology:



The only group that needs clarification is “Omenala”. This is a form of writing that aims at explaining Igbo tradition and culture in the Igbo language. It is simply a genre of its own.

The third stage involved the electronic encoding format. The TEI-xml standard was chosen, as a means of ensuring that the text be put in an easily available format. However, the particular form of the TEI standard that was chosen was TEI-Lite. All the details of the TEI-Standard are not relevant for the work.

Finally, as most of the texts are not available in electronic form, with some only in type-written manuscripts or in worn-out off-set prints, they simply have to be manually

fed into the computer. While this is still going on, effort is also being made to resolve and avoid all possibilities of copyright entanglements.

## **2. POS-Tags**

The main issue here is selecting the appropriate tags from the analysis of the language in Igbo linguistics. A look into Igbo linguistics confirms some level of agreement in terminology, but with little addition arising from my analysis of the language. This is not a major problem, but is simply connected with my effort to make some 'hidden' grammatical issues obvious. However, the tags I have worked out so far can take care of the tagging process at three levels: the clause level, the phrasal level, and the morpheme level. In the next section I shall go into some of the problems that arose in the course of the work, and which still have to be taken care of.

## **3. Matters Arising**

The different problems can be subsumed under the headings of tone marking and corpus development tools.

### **3.1. Problem 1: Tone Marking**

Igbo language data is usually tone marked when analyzed in academic publications. However, compared with written and published Igbo texts from native speakers, such efforts can really be classified as the exception. 99.9% of Igbo texts are not tone marked, because they are written by native speakers and for native speakers. This also applies to the texts collected for the Igbo Corpus Model. To further complicate matters, there are two tone marking systems in Igbo linguistics to choose from. I left the Igbo texts of the Corpus Model unmarked, because any effort at tone marking would mean manually keying in the whole texts! The result of this is that the Igbo Corpus Model can only be of use to native speakers. Fortunately, a new body made up of mainly Igbo linguists has been formed and it is known as the *Igbo Studies Association* (ISA). It has now replaced the former more general body known as the *Society for the Promotion of Igbo Language and Culture* (SPILC). At its inaugural meeting in September 2005 at the University of Nigeria, Nsukka, the new body resolved to take an official stand on the tone marking problem and come up with only one tone marking system for the language. This would be

to the benefit of the language, and shall be taken into consideration in the *Igbo Corpus Model* as soon as the association makes available the official clarification of the issue.

### **3.2. Problem 2: Corpus Tools**

The two main issues here are (1) Corpus linguistics in Nigeria, and (2) Corpus Development Software.

Corpus linguistics is simply an undeveloped field in the Nigerian university system, although efforts are presently being made in various directions to make a change. The major problem that still confronts most departments of linguistics in the country in this regard is finding the trained professionals to start off the development of corpus linguistics.

In terms of Corpus Development software, the main problem has always been whether the program is FULLY Unicode based. Many a corpus manipulation or development software is not fully Unicode based. I have often had the disappointing experience of using a program to put a text in TEI-Lite format, only to find out that the Concordance or KWIC function is not Unicode based and that I cannot even key in any of the Igbo texts with special characters like sub-dotted vowels, available in a Unicode based font like Doulos, Gentium or Code2000 (Uchechukwu 2005). To place a tone mark on such a sub-dotted vowel is like asking for the impossible from most of these programs. And this is the most discouraging part of the work. However, a recent experiment with the *Word Sketch Engine* (<http://www.sketchengine.co.uk/>) has proved very useful and promising. The program can easily handle the sample of Igbo text without tone marks that has been loaded into it. This is indeed a very big relief for me. The next step shall involve examining how the program can handle a tone marked Igbo text.

## **4. CONCLUSION**

As an Igbo saying has it “You do not run away from an unavoidable war simply because you might be shot”. Developing the Igbo corpus is unavoidable work, and it shall not be abandoned because of some of the difficulties encountered. The success of the experiment with the Word Sketch Engine also means that the difficulties can be overcome.

## **Bibliography**

Bendor-Samuel, J.T. (ed.) 1989. *The Niger-Congo languages: a classification and description of Africa's largest language family*. Lanham/New York/London: University Press of America.

Igboanusi, Herbert & Lothar Peter. 2005. *Languages in competition*. Frankfurt am Main: Peter Lang.

Manfredi, Victor/Eze, Ejike (2001/2): 'Igbo'. In Garry, J/Rubino, Carl (Hg.) (2002): *Facts about the world's major languages: an encyclopedia of the world's major languages, past and present*. New York/Dublin: The H.W. Wilson Company. 322 – 330.

Uchechukwu, Chinedu (2004): *The representation of Igbo with the appropriate keyboard*. International Workshop on Igbo Meta-Language, at University of Nigeria, Nsukka, on 18th April, 2004.

Uchechukwu, Chinedu (2005): *The Igbo language and computer linguistics: problems and prospects*.  
[https://www.eurac.edu/Org/LanguageLaw/Multilingualism/Projects/conference2005\\_programme.htm](https://www.eurac.edu/Org/LanguageLaw/Multilingualism/Projects/conference2005_programme.htm)

Williamson, Kay & Roger Blench. 2000. 'Niger-Congo'. In Bernd Heine & Derek Nurse (eds.). *African languages: an introduction*. Cambridge: University Press.