Part-of-speech Tagset and Corpus Development for Igbo, an African Language

Ikechukwu E. Onyenwe
Dept. of Computer Science,
University of Sheffield
Sheffield S1 4DP, UK
i.onyenwe@shef.ac.uk

Dr. Chinedu Uchechukwu
Dept. of Linguistics
Nnamdi Azikiwe University
Anambra State, Nigeria
neduchi@yahoo.com

Dr. Mark Hepple
Dept. of Computer Science,
University of Sheffield
Sheffield S1 4DP, UK
m.r.hepple@shef.ac.uk

Abstract

This project aims to develop linguistic resources to support computational NLP research on the Igbo language. The starting point for this project is the development of a new part-of-speech tagging scheme based on the EAGLES tagset guidelines, adapted to incorporate additional language internal features. The tags are currently being used in a part-of-speech annotation task for the development of POS tagged Igbo corpus. The proposed tagset has 59 tags.

1 Introduction

Supervised machine learning methods in NLP require an adequate amount of training data. The first crucial step for a part-of-speech (POS) tagging system for a language is a well designed, consistent, and complete tagset (Bamba Dione et al., 2010) which must be preceded by a detailed study and analysis of the language. Our tagset was developed from scratch through the study of linguistics and electronic texts in Igbo, using the EAGLES recommendations.

This initial manual annotation is important. Firstly, information dealing with challenging phenomena in a language is expressed in the tagging guideline; secondly, computational POS taggers require annotated text as training data. Even in unsupervised methods, some annotated texts are still required as a benchmark in evaluation. With this in mind, our tagset design follows three main goals: to determine the tagset size, since a smaller granularity provides higher accuracy and less ambiguity (de Pauwy et al., 2012); to use a sizeable scheme to capture the grammatical distinctions at a word level suited for further grammatical analysis, such as parsing; and to deliver good accuracy for automatic tagging, using the manually tagged data. We discuss the development of the tagset and corpus for Igbo. This work is, to the best of our knowledge, the first published work attempting to develop statistical NLP resources for Igbo.

2 Some Grammatical Features of the Igbo Language

2.1 Language family and speakers

The Igbo language has been classified as a Benue-Congo language of the Kwa sub-group of the Niger-Congo family[1] and is one of the three major languages in Nigeria, spoken in the eastern part of Nigeria, with about 36 million speakers[2]. Nigeria is a multilingual country having around 510 living languages[1], but English serves as the official language.

2.2 Phonology

Standard Igbo has eight vowels and thirty consonants. The 8 vowels are divided into two harmony groups that are distinguished on the basis of the Advanced Tongue Root (ATR) phenomenon. They are -ATR: ị [ɪ], ụ [ʊ], a [ɑ], ọ [ɔ] and +ATR: i [i], u [u], e [e], o [o] (Uchechukwu, 2008). Many Igbo words select their vowels from the same harmony group. Also, Igbo is a tonal language. There are three distinct tones

[1] http://nigerianwiki.com/wiki/Languages
[2] http://en.wikipedia.org/wiki/Igbo_people

LAW VIII - The 8th Linguistic Annotation Workshop, pages 93–98,
Dublin, Ireland, August 23-24 2014.

recognized in the language viz; High, Low, and Downstep. The tones are represented as *High* [H] = [́], *Low* [L] = [̀], *downstep* = [̄] (Emenanjo, 1978; Ikekeonwu, 1999) and are placed above the tone bearing units (TBU) of the language.

There are two tone marking systems, either: all high tones are left unmarked and all low tones and downsteps are marked (Green and Igwe, 1963; Emenanjo, 1978), or only contrastive tones are marked (Welmers and Welmers, 1968; Nwachukwu, 1995). We used the first system to illustrate the importance of tonal feature in the language's lexical or grammatical structure. For example, at the lexical level the word *akwa* without a tone mark can be given the equivalent of 'bed/bridge', 'cry', 'cloth', or 'egg'. But these equivalents can be properly distinguished when tone marked, as follows: akwa "cry", akwà "cloth", àkwà "bed or brigde", àkwa "egg". At the grammatical level, an interrogative sentence can be distinguished from a declarative sentence through a change in tone of the person pronouns from a high tone (e.g. ***Ọ*** *nà-àbịa* "He is coming") to a low tone (e.g. ***Ọ̀*** *nà-àbịa* "Is he coming?"). Also, there are syllabic nasal consonants, which are tone bearing units in the language. The nasal consonants always occur before a consonant. For example: ǹdo 'Sorry' or explicitly tone marked as ǹdó.

2.3 Writing System

The Igbo orthography is based on the Standard Igbo by the Ọnwụ Committee (Ọnwụ Committee, 1961). There are 28 consonants: *b gb ch d f g gh gw h j k kw kp l m n nw ny ṅ p r s sh t v w y z*, and 8 vowels (see phonology section). Nine of the consonants are digraphs: *ch, gb, gh, gw, kp, kw, nw, ny, sh*.

Igbo is an agglutinative language in which its lexical categories undergo affixation, especially the verbs, to form a lexical unit. For example, the word form *erichaṛiṛị* is a verbal structure with four morphemes: verbal vowel prefix *e-*, verb root *-ri-*, extensional suffix *-cha-*, and a second extensional suffix *-ṛiṛị*. Its occurrence in the sentence "*Obi must eat up that food*" is *Obi ga-erichaṛiṛị nri ahụ*, that is, *Obi aux-eat.completely.must food DET*. Igbo word order is Subject-Verb-Object (SVO), with a complement to the right of the head.

2.4 Grammatical Classes

Generally, Emenanjo (1978) identified the following broad word classes for Igbo: verbal, nominal, nominal modifier, conjunction, preposition, suffixes, and enclitics. The verbal is made up of verbs, auxiliaries and participles, while the nominal is made up of nouns, numerals, pronouns and interrogatives. Nouns are further classified into five lexical classes, viz; proper, common, qualificative, adverbial and ideophones. However, we identified extra five in the tagset design phase (see the appendix). Nominal modifiers occur in a noun phrase. Its four classes are adjectives, demonstratives, quantifiers and pronominal modifiers. Conjunctions link words or sentences together, while prepositions are found preceding nominals and verbals and cannot be found in isolation. Suffixes and enclitics are the only bound elements in the language. Suffixes are primarily affixed to verbals only, while enclitics are used with both verbals and other word classes. Suffixes are found in verb phrase slots and enclitics can be found in both verb phrase and noun phrase slots. The language does not have a grammatical gender system.

3 Language Resources

The development of NLP resources for any language is based on the linguistics resources available for the language. This includes appropriate fonts and text processing software as well as the available electronic texts for the work. The font and software problems of the language have been addressed through the Unicode development (Uchechukwu, 2005; Uchechukwu, 2006). The next is the availability of Igbo texts.

Any effort towards the Igbo corpus development is a non-trivial task. There are basic issues connected with the nature of the language. The first major surprise is that Igbo texts 'by native speakers' written 'for native speakers' vary in forms due to dialectal difference and are usually not tone-marked. Indeed, the tone marking used in the sections above are usually found in academic articles. It would be strange to find an Igbo text (literary work) that is fully tone marked and no effort has been made to undertake a tone marking of existing Igbo texts. Such an effort looks impossible as more Igbo texts are written and

published. Such is the situation that confronts any effort to develop an Igbo corpus. Hence, developing NLP resources for the language has to start with the available resources; otherwise, such an endeavour would have to first take a backward step of tone marking all the texts to be added to its corpus and normalizing the dialectal differences. This is a no mean task.

It is for this reason that we chose the New World Translation (NWT) Bible version for Igbo corpus with its English parallel text[3]. The NWT Bible does not adopt a particular tone marking system, neither is there a consistent use of tone marks for all the sentences in the Bible. Instead, there is narrow use of tone marks in specific and restricted circumstances throughout the book. An example is when there is a need to disambiguate a particular word. For instance, *ihe* without tone mark could mean 'thing' or 'light'. These two are always tone marked in the Bible to avoid confusion; hence *ìhè* 'light' and *íhé* 'thing'. The same applies to many other lexical items. Another instance is the placement of a low tone on the person pronouns to indicate the onset of an interrogative sentence, which otherwise would be read as a declarative sentence. This particular example has already been cited as one of the uses of tone mark in the language. Apart from such instances, the sentences in the Bible are not tone marked. As such, one cannot rely on such restricted use of tone marks for any major conclusions on the grammar of the language. With regard to corpus work in general, the Bible has been described as consistent in its orthography, most easily accessible, carefully translated (most translators believe it is the word of God), and well structured (books, chapters, verses), etc. (Resnik et al., 1999; Kanungo and Resnik, 1999; Chew et al., 2006). The NWT Bible is generally written in standard Igbo.

4 Tokenization

We outline here the method we used in the tokenization of the text. For the sake of a start-up, we tokenized based on the whitespace. The Igbo language uses whitespace to represent lexical boundaries; we used the following regex:

Separate characters if the string matches:

- "ga-" or "n"' or "N"' or "na-" or "Na-" or "ana-" or "ịna-"; for example, the following samples *n'elu, na–erugharị, ịna-akwa, ana-egbu* in the Bible will be separated into *n', elu, na–, erugharị, ịna-, akwa, ana-, egbu* tokens.
- Any non-zero length sequence consisting of a–z, A–Z, 0–9, combining grave accent (̀), combining acute accent (́), combining dot below (̣); for example, these words *ìhè, ahụ́, ájá* in the corpus will be separated as tokens with their diacritics.
- Any single character from: left double-quotation mark ("), right double-quotation mark (''), comma (,), colon (:), semicolon (;), exclamation (!), question (?), dot (.).
- Any single non-whitespace character.

In place of sentence splitting, we use verses since all 66 books of the Bible is written in verse level. Our major aim is to use this Igbo corpus to implement our new tagset, which will capture all the inflected and non-inflected tokens in the corpus. For lack of space, issues with tokenization with respect to morphemes, manual annotation implemetations and platform used will not be discussed in this paper.

5 Tagset Design

We adopt the (Leech, 1997) definition of a POS tagset as a set of word categories to be applied to the tokens of a text. We designed our tagset following the standard EAGLES guidelines, diverging where necessary (e.g. EAGLES, which favours European languages, specifies *articles* at the obligatory level, but this category does not apply for Igbo). A crucial question in tagset design is the extent of fine-grained distinctions to encode within the tagset. A too coarsely grained tagset may fail to capture distinctions that would be valuable for subsequent analysis, e.g. syntactic parsing; too fine-grained may make automatic (and manual) POS tagging difficult, resulting in errors that lead to different problems for later processing. In what follows, we introduce a sizeable tagset granularity with the intention of providing a basis for practical POS tagging.

[3]Obtained from `jw.org`.

NNM	Number marking nouns	NNT	Instrumental nouns
NNQ	Qualificative nouns	VrV	$-rV$ implies suffix
NND	Adverbial nouns	VCJ	Conjunctional verbs
NNH	Inherent complement nouns	α_XS	any POS tag with affixes
NNA	Agentive nouns		

Table 1: Selected distinctive tags from the tagset scheme

The tagset is intended to strike an appropriate balance for practical purposes regarding granularity, capturing what we believe will be the key lexico-grammatical distinctions of value for subsequent processing, such as parsing. Further subcategorization of the grammatical classes, as described in section 2.4, results in 59 tags which apply to whole tokens (produced by the tokenisation stage described above). An important challenge comes from the complex morphological behaviour of Igbo. Thus, a verb such as *bịa*, which we assign the tag VSI (a verb in its simple or base form), can combine with extensional suffixes, such as *ghị* and *kwa*, to produce variants such as *bịaghị*, *bịakwa* and *bịaghịkwa*, which exhibit similar grammatical behaviour to the base form. As such, we might have assigned these variants the VSI tag also, but have instead chosen to assign VSI_XS, which serves to indicate both the core grammatical behaviour and the presence of extensional suffixes. In *abịakwa*, we find the same base form *bịa*, plus a verbal vowel prefix *a*, resulting in the verb being a participle, which we assign the tag VPP_XS. For the benefit of cross-lingual training and other NLP tasks, a smaller tagset that captures only the grammatical distinctions between major classes is required. The present 59 tags can easily be simplified to a coarse-grained tagset of 15 tags, which will principally preserve just the core distinctions between word classes, such as nouns, verb, adjective, etc.

Athough Emenanjo (1978) classified **ideophones** as a form of noun, we have assigned them a separate tag **IDEO**, as these items can be found performing many grammatical functions. For instance, the **ideophone** *kọị*, "to say that someone walks *kọị kọị*" has no nominal meaning, rather its function here is adverbial. A full enumeration of this scheme is given in the appendix.

5.1 The developement of an POS tagged Igbo Corpus

Here we analyse the manual POS tagging process that is ongoing based on the tagset scheme. The Bible books were allocated randomly to six groups, producing six corpora portions of approximately 45,000 tokens each. Our plan was for each human annotator to tag at least 1000 tokens per day, resulting in complete POS tagging in 45 days. The overall corpus size allocated is 264,795 tokens of the new testament Bible. There are six human annotators, who are students of the Department of Linguistics at Nnamdi Azikiwe University, Awka, supervised by a senior lecturer in the same department; giving an effective total of seven human annotators. Additionally, a common portion of the corpus (38,093 tokens) was given to all the annotators, as a basis for calculating inter-annotator agreement.

6 Conclusions

We have outlined our current progress in the development of a POS tagging scheme for Igbo from scratch. Our project aims to build linguistic computational resources to support research in natural language processing (NLP) for Igbo. It is important to note that these tags are applicable on unmarked, not fully marked, and fully tone marked Igbo texts, since the fully tone marked tokens play the same grammatical roles as in the none tone marked texts, written by native speakers for fellow native speakers.

Our method of tagset design could be used for other African or under-resourced languages. African languages are morphologically rich, and of around 2000 languages in the continent, only a small number have featured in NLP research.

Acknowledgements

We acknowledge the support of Tertiary Education Trust Fund (TETFund) in Nigeria, and would like to thank Mark Tice for his useful comments and help in preparing this paper.

References

Cheikh M. Bamba Dione, Jonas Kuhn, and Sina Zarrieß. 2010. Design and Development of Part-of-Speech-Tagging Resources for Wolof (Niger-Congo, spoken in Senegal). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. ELRA).

Peter A. Chew, Steve J. Verzi, Travis L. Bauer, and Jonathan T. McClain. 2006. Evaluation of the Bible as a Resource for Cross-Language Information Retrieval. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*. Association for Computational Linguistics.

E. Nọlue Emenanjo. 1978. *Elements of Modern Igbo Grammar: A Descriptive Approach*. Ibadan Ox. Uni. Press.

Margaret M. Green and G. Egemba Igwe. 1963. *A descriptive grammar of Igbo*. London: Oxford University Press and Berlin: Akademie-Verlag.

Clara Ikekeonwu. 1999. *"Igbo", Handbook of the International Phonetic Association*. C. U. Press.

Tapas Kanungo and Philip Resnik. 1999. The Bible, Truth, and Multilingual OCR Evaluation. In *Proceedings of SPIE Conf. on Document Recognition and Retrieval*, pages 86–96.

Geoffrey Leech. 1997. *Introducing Corpus Annotation*. Longman, London.

P. Akujuoobi Nwachukwu. 1995. Tone in Igbo Syntax. Technical report, Nsukka: Igbo Language Association.

Guy de Pauwy, Gilles-Maurice de Schryverz, and Janneke van de Loo. 2012. Resource-Light Bantu Part-of-Speech Tagging. In *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages*, pages 85–92.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'. *Computers and the Humanities*, 33.

Chinedu Uchechukwu. 2005. The Representation of Igbo with the Appropriate Keyboard. In Clara Ikekeonwu and Inno Nwadike, editors, *Igbo Lang. Dev.: The Metalanguage Perspective*, pages 26–38. CIDJAP Enugu.

Chinedu Uchechukwu. 2006. Igbo Language and Computer Linguistics: Problems and Prospects. In *Proceedings of the Lesser Used Languages and Computer Linguistics Conference*. European Academy (EURAC).

Chinedu Uchechukwu. 2008. African Language Data Processing: The Example of the Igbo Language. In *10th International pragmatics conference, Data processing in African languages*.

Beatrice F. Welmers and William E. Welmers. 1968. Igbo: A Learner's Manual. Published by authors.

Ọnwụ Committee. 1961. The Official Igbo Orthography.

A A Tagset Design for the Igbo Language

Noun Class	
Tag	Description/Example
NNP	**Noun Proper**. *Chineke* 'God', Onyeka, Okonkwo, Osita.
NNC	**Noun Common**. *Okụ* 'fire', *ụwa* 'earth', *osisi* 'tree, stick', *ala* 'ground', *eluigwe* 'sky, heaven'
NNM	**Number Marking Noun**. *Ndị* 'people', *nwa* 'child', *ụmụ* 'children'. *ndị* is classified as a common noun with an attached phrase of "thing/person associated with" (Emenanjo, 1978). *ndị* preceding a noun marks plurality of that noun, *nwa* marks it singular (e.g. *nwa agbọghọ* 'a maiden'), and *ụmụ* also indicate plurality (e.g. *ụmụ agbọghọ* 'maidens').
NNQ	**Qualificative noun**. Nouns that are inherently semantically descriptive. E.g. *ogologo* [height, long, tall]
NND	**Adverbial noun**. This lexical class function to modify verbals, e.g. *O ji **nwayọọ** eri nri ya*
NNH	**Inherent Complement**. Igbo verb has a [verb + NP/PP] structure. NP/PP are the verb complement. They cooccur with the verb, at times quite distant from the verb, e.g. (1) *ịgụ egwu* 'to sing' , (2) *iti ịgba* 'to drum', (3) *igwu ji* 'harvest yam'.
NNA	**Agentive Noun**. Nouns are formed through verbs nominalization. Compare (1) with *ọgọ egwu* 'singer' and (2) with *oti ịgba* 'drummer'. For links *NNAV . . . NNAC*.
NNT	**Instrumental Noun**. Refer to instruments and are formed via nominalization. Compare (3) with *ngwu ji* 'digger'. For links *NNTV . . . NNTC*.
NOTE: We introduced link indicators in NNA and NNT, V and C, Where V and C stand for verbal and Complementary respectively. So, NNAV indicates derivation from the verbal component of the inherent complement verb and NNAC is the inherent complement of the whole verbal complex. E.g., *ọgụ/NNAV egwu/NNAC*. Also, NNTV and NNTC, where NNTV is derived from the verbal component of the inherent complement verb and NNTC is the inherent complement of the whole verbal complex. E.g. , *ngwu/NNTV ji/NNTC*	

Verb Class	
VIF	**Infinitive**. Marked through the addition of the vowel [i] or [ị] to the verb root.
VSI	**Simple verb**. Has only one verb root.
VCO	**Compound Verb**. Involves a combination of two verb roots.
VIC	**Inherent Complement Verb (ICV)**. Involves the combination of a simple or compound verb with a noun phrase or a prepositional phrase. It gives rise to the structures (1) V + NP, or (2) V + PP
VMO	**Modal Verb**. Its formed by inherent complement verbs and simple verbs. [See the section on suffixes]
VAX	**Auxiliary Verb**. ga [Future marking], na [progressive]
VPP	**Participle**. Always occurs after the auxiliary, and prefixed e/a to the verb root using vowel harmony.
VCJ	**Conjunctional Verb**. A verb that has a conjuntional meaning, especially in narratives: *wee*
VBC (BVC)	**Bound Verb Complment or Bound Cognate Noun**. Its formed by harmonizing prefix *a/e* to the verb root. It looks like the participle but occurs after the participle in same sentence as the verb. It can be formed from every verb.
VGD	**Gerund**. Reduplication of the verb root plus harmonizing vowel o/ọ. Also, internal vowel changes can occur. E.g. *ba* 'enter' [ọ + bụ + ba]=*ọbụba* 'the entering'
Inflectional Class	
VrV	$-rV$ (e.g. *-ra*). If attached to an active verb, it means simple past; but a stative meaning with a stative verb.
VPERF	**Perfect** (e.g. *-la/-le*, *-go*). Describes the 'perfect tense'. *-la/-le* obeys vowel harmony and the variant *-go* does not.
Other part-of-speech tags	
ADJ	**Adjective**. The traditional part of speech 'adjective' that qualifies a noun. Igbo has very few of them.
PRN	**Pronoun**. The 3 persons are 1st (sing + pl), 2nd (sing + pl), and 3rd (sing + pl) person Pronouns.
PRNREF	**Reflexive Pronoun**. Formed by combination of the personal pronouns with the noun *onwe* 'self'.
PRNEMP	**Emphatic pronoun**. This involves the structure [pronoun+onwe+pronoun].
ADV	**Adverb**. Changes or simplifies the meaning of a verb. They are few in Igbo.
CJN	**Conjunction**. There are complex and simple conjunctions distinguish based on grammatical functions viz; co-rodinators, sub-ordinators and correlatives. Link indicators *CJN1...CJN2* are for "correlative CJN". E.g. *ma/CJN1...ma/CJN2*.
PREP	**Preposition**. The preposition *na* is realised as *n'* if the modified word begins with a vowel.
WH	**Interrogative**. Questions that return useful data through explanation. *Ònye, gịnị, olee, ...*
PRNYNQ	**Pronoun question**. Questions that return YES or NO answer. E.g. *m̀, à, hà, ò, ̀ọ, ...*
IDEO	**Ideophone**. This is used for sound-symbolic realization of various lexico-grammatical function. E.g. *nịganịga, mụrịị, kọị, etc.*
QTF	**Quantifier**. This can be found after their nominals in the NP structure. E.g. *dum, naabọ, nille*.
DEM	**Demonstrative**. This is made up of only two deictics and always used after their nominals. E.g. *a, ahụ*.
INTJ	**Interjection**.*Ee*
FW	**Borrowed word**. *amen*.
SYM	**Punctuation**. It includes all symbols.
CD	**Number**. This includes all digits 1,2,3, ... and *otu, mbụ, abụa, atọ, ...*
DIGR	**Digraph**. All combined graphemes that represent a character in Igbo, which occur in the text. *gb, gw, kp, nw, ...*
TTL	**Title** . Includes foreign and Igbo titles. E.g. *Maazị*.
CURN	**Currency**.
ABBR	**Abbreviation**.
Any type of suffixes	
α_XS	**any POS tag with affixes**. for $\alpha \in$ {VIF, VSI, VCO, VPP, VGD, VAX, CJN, WH, VPERF, VrV, PREP, DEM, QTF, ADJ,ADV}. See verb, other POS, inflectional classes.
NOTE: Tags with affixes identify inflected token forms in the corpus for use in further analysis, e.g. morphology. For practical POS tagging, such tags may be simplified, i.e. $\alpha\_XS \Rightarrow \alpha$.	
Any type of Enclitics	
ENC	**Collective**. *cha, sị nụ, kọ* – means all, totality forming a whole or aggregate. **Negative Interrogative**. *dị, rị, dụ* – indicates scorn or disrespect and are mainly used in Rhetorical Interrogatives. **Adverbial 'Immediate present and past'**. *fọ/hụ* – it indicates action that is just/has just taking/taken place. *rịị* – indicates that an action/event has long taken place **Adverbial 'Additive'**. *kwa (kwọ), kwu* – mean 'also', 'in addition to', 'denoting', 'repetition or emphasis'. **Adverbial 'Confirmative'**. *nọọ (nọọ; nnọọ)* – this means really or quite.

B The Major Classes of the Tagset

ADJ	adjective	FW	foreign word	QTF	quantifier	ADV	adverb	NNC	common noun
INTJ	interjection	SYM	symbol	CJN	conjunction	NNP	proper noun	PREP	preposition
WH	interrogative	PRN	pronoun	V	verb	CD	number	DEM	demonstration
There is no **article** in the language.									