Resources and Tools for Automated Speech Segmentation of the African Language Naija (Nigerian Pidgin)

Brigitte Bigi[1(✉)], Oyelere S. Abiola[2], and Bernard Caron[3]

[1] LPL, CNRS, Aix-Marseille Univ.,
5, Avenue Pasteur, 13100 Aix-en-Provence, France
brigitte.bigi@lpl-aix.fr
[2] UMR 7114, MODYCO-CNRS, Université Paris Nanterre,
200 Avenue de la Republique, 92000 Nanterre, France
biolaoye2@gmail.com, oyelere_sa@parisnanterre.fr
[3] Llacan, CNRS, Inalco, Paris, France
bernard.caron@cnrs.fr
http://www.lpl-aix.fr/~bigi

Abstract. The development of HLT tools inevitably involves the need for language resources. However, only a handful number of languages possesses such resources. This paper presents the development of HLT tools for the African language Naija (Nigerian Pidgin), spoken in Nigeria. Particularly, this paper is focusing on developing language resources for a tokenizer, an automatic speech system for predicting the pronunciation of the words and their segmentation.

The newly created resources are integrated into SPPAS software tool and distributed under the terms of public licenses.

Keywords: Speech · Segmentation · Naija · HLT

1 Introduction

The development of Human Language Technologies (HLT) tools is a way to break down language barriers. There are approximately 6000 languages in the world but unfortunately, only a handful possess the linguistic resources required for implementing HLT technologies [1]. Large corpora datasets for most of the under-resourced languages are created by HLT researchers for Natural Language Processing (NLP) or for Speech Technologies. African languages form about 30% of the world languages and their native speakers form 13% of the world population [9]. However, "NLP in Africa is still in its infancy; of about 2000 languages, a very few have featured in NLP research and resources, which are not easily found online." [10]. With a 182M population, Nigeria counts about 92M Internet users for June, 2015, i.e. 51.1% of its population[1] and it's constantly

[1] Source: http://www.internetworldstats.com/af/ng.htm - 2017-10.

Z. Vetulani et al. (Eds.): LTC 2017, LNAI 12598, pp. 164–173, 2020.
https://doi.org/10.1007/978-3-030-66527-2_12

growing. The official language is English but over 527 individual languages are spoken in Nigeria[2]. Recently, the Igbo language was one of the Nigerian languages investigated for NLP [10].

Among these Nigerian languages, Nigerian Pidgin English (NPE) is a post-creole continuum that is spoken as a first language (L1) by 5 million people, while over 70 million people use it as a second language (L2) or as an inter-ethnic means of communication in Nigeria and in Nigerian Diaspora communities. The origin of NPE is generally described as a development out of an English-lexified jargon attested in the 18th Century in the coastal area of the Niger delta (River State), with some lexical influence from Krio through the activities of missionaries from Sierra Leone [6]. The heartland of NPE is the Niger Delta, with Lagos and Calabar as secondary extensions. NPE is identified by "pcm" in the iso-639-3 language codes.

Since the independence of Nigeria in 1960, NPE has been rapidly expanding from its original niche in the Niger delta area, to cover two-thirds of the country, up to Kaduna and Jos, and is now deeply rooted in the vast Lagos conurbation of over 20 million people. It has become, over the last 30 years the most important, most widely spread, and perhaps the most ethnically neutral lingua franca used in the country today. In this geographical expansion, and as it conquers new functions (e.g. in business, on higher education campuses, in the media and in popular arts), NPE is subject to extensive contact and influence from its original lexifier, i.e. English and from the multitude of vernacular Nigerian languages. A mixed language has emerged that is fast expanding (both in geography and function) and rapidly changing. The name **Naija** (meaning 'Nigeria' in NPE) is used to describe this language learnt and used as an L2 in most of Nigeria, and differenciate it from the creolised variety (NPE) spoken as an L1 in the Niger delta (see [5] for a short characterization of Naija). Naija is the object of this paper on the development of HLT as part of the NaijaSynCor project[3]. It aims at describing the language in its geographical and sociological variations, based on a 500k word corpus annotated and analyzed with cutting edge HLP tools developed for corpus analysis.

Then, the development of speech technologies for Naija faces the following problems: (1) lack of language resources (lexicon, corpora, ...), not to mention digital resources; and (2) acoustic and phonological characteristics that still need to be properly investigated. These issues are shared with most under-resourced languages, and linguists are currently looking for solutions to solve or to avoid them. Nevertheless, language data collection is still a challenging and fastidious task.

This paper describes the development of a corpus and some language resources for Naija as part of a corpus-based project, which aims at evaluating the nativization of the language. Such newly created linguistic resources were integrated into SPPAS software tool [2] for a tokenizer, an automatic speech system for predicting the pronunciation of words and their segmentation. This

[2] Source: https://www.ethnologue.com/country/NG - 2017-06.
[3] http://naijasyncor.huma-num.fr/.

paper describes such resources at two stages of the process: at the beginning of the project and at its end.

2 Corpus Creation

For the NaijaSynCor project, a total of 384 samples of oral corpus (monologues, dialogues), an 6 min each, is to be collected from 380 speakers so as to represent the widest scope of functions and locations of Naija in the country. The speech recordings are done using professional digital recorders and wireless microphones - one per speaker.

At the initial stage of the project, 8 of these recordings were partially manually transcribed and time-aligned at the phonetic level (Table 1). The transcriptions use the Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA) code, a machine-readable phonetic alphabet that was originally developed by [12]. The recordings are totalling 3 min 29 s in length, 4 men (M) and 4 women (W). Only these files were available to construct our first HLT tools; recordings were being collected and their orthographic transcription was done gradually during the project.

Table 1. Description of the transcribed corpus, manually time-aligned at the phoneme level.

File (wav)	Recording duration (in sec.)	Speech duration (in sec.)	Nb of phon-	Speech rate (phon/sec)
M_1	32.578	20.817	254	12.20
M_2	35.155	23.509	281	11.95
M_3	48.431	35.960	403	11.20
M_4	40.243	20.213	233	11.53
W_1	33.698	28.708	360	12.54
W_2	35.926	28.174	258	9.16
W_3	37.311	27.790	284	10.22
W_4	34.239	24.087	263	10.92

At the end of the project, 80 files representing about 8 h of recordings were manually annotated:

1. orthographic transcription time-aligned into Inter-Pausal Units;
2. time-aligned tokens;
3. X-SAMPA transcription time-aligned at the syllables level.

Orthographic transcription is often the minimum requirement for a speech corpus, as it is the entry point for most HLT tools. Corpora are using the orthography developed by [4] in her work on Lagos Nigerian Pidgin. This etymological

orthography - adapted from the lexifier language orthography, i.e. English - has been chosen preferably to the phonological script used by linguists as it is spontaneously used by educated Nigerians, and thus easier to teach to transcribers. Code-switched to English sections are identified by dedicated boundaries.

The following are examples of transcribed speech:

(W_2) '*so, all Edo people wey don travel go different-different-different places, everybody go come travel come back.*' So, all the Edo people who have travelled far and wide, everybody will return home.
(M_2) '*So, we don carry di matter come again, as we dey always carry am come.*' So, we have brought the topic again, as we always bring it.

3 Phonetic Description of Naija

At the beginning of the project, the phonetic transcription of a few minutes of speech enabled us to establish the list of the phonemes that are mostly used by the speakers. While the list of consonants is pretty close to the English one, the list of vowels used in Naija language is very different.

As shown in Table 2, only the sounds /dZ/ and /tS/ were observed infrequently, and the English /D/ and /T/ were not observed in the initial manual phonetic transcription but in the whole corpus. The other consonants of English are shared by both languages except /Z/.

But as also shown in the Table 2, the set of vowels Naija and English are sharing is only: /E/, /i/, /u/ and diphtongs. Six different nasalized vowels were observed in the corpus, but with a small number of occurrences.

Table 2. Phonemes inventory and occurrences in the manually time-aligned files and in the whole corpus

b	d	g	k	p	t	tS	dZ		m	n	N	j	w	
37	163	52	90	54	119	3	3		98	140	4	26	89	
5248	16549	5715	8361	5069	12062	970	1483		8173	11361	122	5539	7790	
l	r\	S	f	s	z	v	h	T	D		OI	aI	aU	eI
57	58	23	48	141	12	24	12	0	0		0	37	8	0
5647	5643	1581	6210	13991	1780	2454	1105	32	113		352	5421	1100	151
a	e	E	i	o	O	u	a∼	e∼	E∼	i∼	O∼	u∼		
203	123	111	221	93	126	74	21	1	18	18	20	3		
17141	12348	9601	22368	9110	14088	8195	2486	431	1983	2964	3901	385		

4 Creating Resources for HLT Tools

4.1 Vocabulary

A lexicon was created with both all English words and specific words observed in the corpus. In a first version of the lexicon, established in 2017, we added

about 700 words. Gradually, as more corpora were transcribed, new words with their orthographic variants were added.

At the end of the project, the orthographic transcriptions 8 h recordings of the corpus contains 90k words. They represent a vocabulary of 4,600 different words; and among them 1,540 (33%) are specific to Naija language: there are not in the English vocabulary. Among the 4,600 words of the vocabulary, 2,040 (44%) are occurring only once. The 10 most frequent words are covering 24% of the pronounced words:

1. dey: 3658 occurrences
2. go: 2794 occurrences
3. i: 2697 occurrences
4. di: 2554 occurrences
5. you: 2168 occurrences
6. na: 1745 occurrences
7. for: 1634 occurrences
8. sey: 1597 occurrences
9. e: 1497 occurrences
10. no: 1477 occurrences

4.2 Pronunciation Dictionary

A pronunciation dictionary was manually created including observed words only. The dictionary was originally created by extracting the lexicon of the corpus published in annex of [4]. The observed pronunciations of the corpus of Table 1 were added to the dictionary. Created en 2017, the first dictionary was made of 4.7k pronunciations of 3.7k words.

At the end of the project, the pronunciations observed in the whole corpus were automatically extracted and added to the dictionary. The final dictionary contains 10.6k pronunciations of 5.7k words. Here is an extract of its content:

```
above [above] a b O f
above(2) [above] a b o f
above(3) [above] a b o v
abroad [abroad] a b r\ O
abroad(2) [abroad] a b r\ O d
figures [figures] f i g O
file [file] f aI l
fill [fill] f i
fill(2) [fill] f i l
```

4.3 Acoustic Model

Acoustic models were created using the HTK Toolkit [13], version 3.4. The models are Hidden Markov models (HMMs). Typically, HMM states are modeled by Gaussian mixture densities whose parameters are estimated using an expectation maximization procedure. Acoustic models were trained from 16 bits, 16,000

HZ wav files. The Mel-frequency cepstrum coefficients (MFCC) along with their first and second derivatives were extracted from the speech in the standard way (MFCC_D_N_Z_0). The training procedure is based on the VoxForge tutorial. The outcome of this training procedure is dependent on both: 1/ the availability of accurately annotated data; and 2/ on good initialization.

Of course, such requirements are difficult to fit in, particularly for under-resourced languages. The initialization of the models creates a prototype for each phoneme using time-aligned data. In the specific context of this study particularly at the beginning of the project with a lack of training data, this training stage has been switched off. It has been replaced by the use of phoneme prototypes already available in some other languages. The articulatory representations of phonemes are so similar across languages that phonemes can be considered as units which are independent from the underlying language [11]. In SPPAS package, nine acoustic models of the same type - i.e. same HMMs definition and same MFCC parameters, were freely distributed with a public license so that the phoneme prototypes can be extracted and reused: English, French, Italian, Spanish, Catalan, Polish, Mandarin Chinese, Southern Min.

To create an initial model for Naija language, most of the prototypes of English language were used. The others were extracted from French language in majority then Southern Min (3 of the nasals), Italian (3 sounds) Polish (/O/) and Spanish (/T/).

The following fillers were also added to the model in order to be automatically time-aligned too: silence, noise, laughter.

This approach enables the acoustic model to be trained by a small amount of target language speech data [7]. The initial Naija model was created by using the 8 files described in Table 1 only. At the end of the project, the whole corpus was introduced in the training procedure and an updated model was created.

5 HLT Tools

In recent years, the SPPAS software tool has been developed to automatically produce annotations, including the alignment of recorded speech sounds with its phonetic annotation. The multi-lingual approaches that are proposed enabled us to adapt some of the automatic annotations of SPPAS to Naija language. An example of Text Normalization, Phonetization and Alignment of a Naija speech segment is proposed in Fig. 1.

5.1 Automatic Tokenization and Phonetization

Tokenization of the Naija language is very similar to the English one. For the purpose of our multimodal studies, we slightly adapted the Text Normalization [1] and Phonetization systems [3]. For the text normalization, we had only to add the list of words of the Naija language into the “resources” folder of SPPAS. From the orthographic transcription, the text normalization system produces tokens (first line of annotations in Fig. 1). These can then be used by the automatic

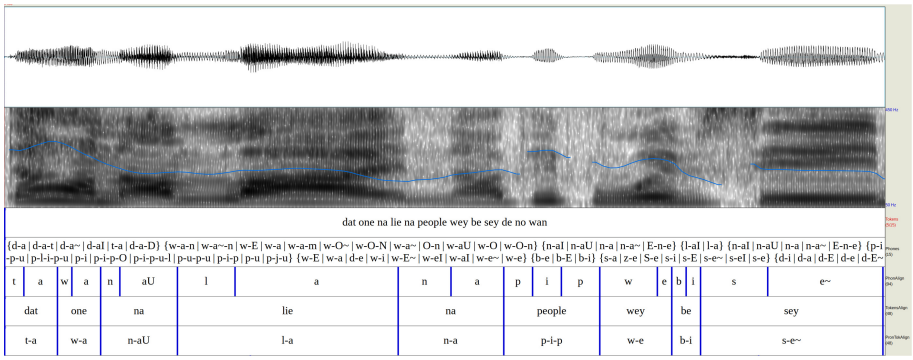

Fig. 1. Example of result of the automatic annotations of Naija

phonetization system (second line of Fig. 1). For that purpose, we simply copied the pronunciation dictionary of Naija into the "resource" folder of SPPAS.

5.2 Automatic Alignment

Forced-alignment is the task of automatically positioning a sequence of phonemes in relation to a corresponding continuous speech signal. Given a speech utterance along with its phonetic representation, the goal is to generate a time-alignment between the speech signal and the phonetic representation. Experiments in this paper were carried out by adding the Naija acoustic model into the "resources" folder of SPPAS. The automatic alignment can be carried out either using HTK (HVite) or Julius decoder engines [8]. Julius is the default aligner used in SPPAS. It produced the alignment of phonemes and tokens as shown in the 3rd and 4th lines of Fig. 1.

5.3 Experiments

Some experiments were conducted to evaluate the accuracy of the phoneme alignments. It was evaluated using the Unit Boundary Positioning Accuracy - UBPA that consists in the evaluation of the delta-times (in percentage) comparing manual phonemes boundaries with the automatically aligned ones. Obviously, the main acoustic model can't be evaluated because all the available data was used to train the model. However, we performed some experiments to have a quick glance at the accuracy of the alignments.

An initialization model was created only from the prototypes already available in the other languages, i.e. without using any Naija data nor training procedure. UBPA of such model is 88.57% in a delta-time of 40 ms. This first result confirms the suitability of a cross-lingual approach to create an acoustic model for the speech segmentation task, at least to create an initial one.

The other experiments were performed using the leave-one-out algorithm: 8 models were created. Each model was trained on 7 of our files, and the model

Table 3. UBPA of Naija automatic alignment

Delta T(automatic)-T(manual)	Initial model		Final model	
	Count	Percent	Count	Percent
$-0.030 \leq$ Delta < -0.040	39	1.67%	56	2.40%
$-0.020 \leq$ Delta < -0.030	91	3.90%	97	4.15%
$-0.010 \leq$ Delta < -0.020	222	9.50%	207	8.86%
$0 \leq$ Delta < -0.010	616	26.37%	622	26.63%
$0 <$ Delta $< +0.010$	550	23.55%	579	24.79%
$+0.010 \leq$ Delta $< +0.020$	401	17.17%	441	18.88%
$+0.020 \leq$ Delta $< +0.030$	169	7.24%	144	6.16%
$+0.030 \leq$ Delta $< +0.040$	55	2.35%	52	2.23%
UBPA	91.35%		94.09%	

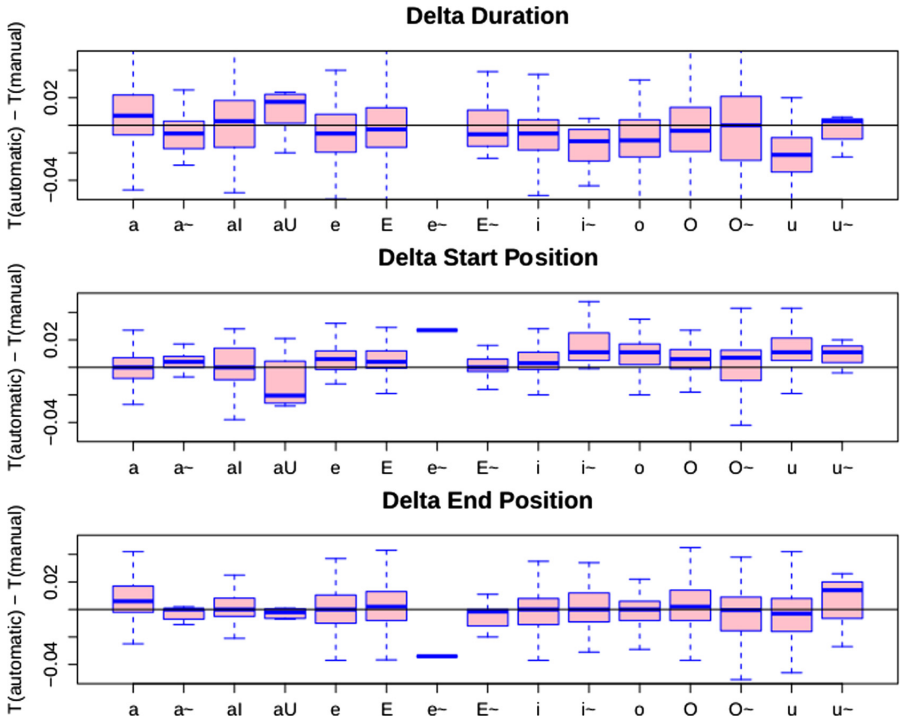

Fig. 2. Detailed results of Naija automatic alignment of vowels

was used to time-align the remaining file. The resulting UBPA is then 91.35%, with a detailed result in Table 3. Of course, introducing Naija manually created data into the training procedure increased significantly the accuracy, even if such

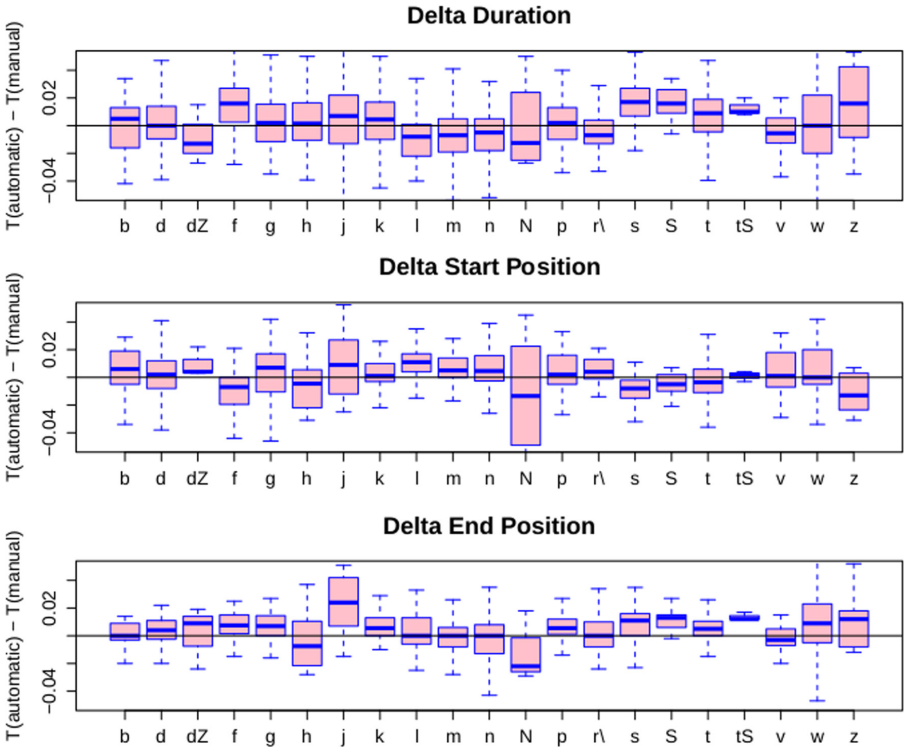

Fig. 3. Detailed results of Naija automatic alignment of consonants

data are only 3 min 29 s long. Finally, at the end of the project, accuracy of the model was enhanced with the whole data corpus as shown in Table 3.

UBPA is a unique measurement suitable to get a quick idea of the accuracy of a model or to compare the quality of several models. However, phoneticians often prefer a qualitative evaluation, as we propose in Fig. 3 and 2 for the final model. We can observe that the automatic system is slightly reducing the duration of the vowels except for /a/, and /aU/, mainly because the beginning of the vowels occurs later that the expected one.

6 Conclusion

This paper presented the first linguistic resources for the Naija language. It is shown that they are useful for HLT tools: it made Text Normalization (including a tokenizer), Phonetization and Alignment automatic annotations available for Naija. These resources were gradually improved and updated as the project progresses. The lexicon, the pronunciation dictionary and the acoustic model are all freely distributed into SPPAS since version 1.9 for the initial model and version 3.0 for the final model.

Acknowledgements. This work was financed by the French "Agence Nationale pour la Recherche" (ANR-16-CE27-0007).

References

1. Bigi, B.: A multilingual text normalization approach. In: Vetulani, Z., Mariani, J. (eds.) LTC 2011. LNCS (LNAI), vol. 8387, pp. 515–526. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-08958-4_42
2. Bigi, B.: SPPAS - multi-lingual approaches to the automatic annotation of speech. Phonetician **111–112**, 54–69 (2015). http://www.isphs.org/Phonetician/Phonetician_111-112.pdf#page=54
3. Bigi, B.: A phonetization approach for the forced-alignment task in SPPAS. In: Vetulani, Z., Uszkoreit, H., Kubis, M. (eds.) LTC 2013. LNCS (LNAI), vol. 9561, pp. 397–410. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43808-5_30
4. Deuber, D.: Nigerian Pidgin in Lagos: Language Contact, Variation and Change in an African Urban Setting. Battlebridge Publications (2005)
5. Esizimetor, D., Egbokhare, F.: Naija. Hawai University Web Site. Language Varieties, 4 July 2014. http://www.hawaii.edu/satocenter/langnet/definitions/naija.html
6. Faraclas, N.: A Grammar of Nigerian Pidgin. Ph.D. thesis, Berkeley University of California (1989)
7. Le, V., Besacier, L., Seng, S., Bigi, B., Do, T.: Recent advances in automatic speech recognition for Vietnamese. In: International Workshop on Spoken Languages Technologies for Under-resourced languages, pp. 47–52. Hanoi, Vietnam (2008). http://www.lpl-aix.fr/~bigi/Doc/le2008sltu.pdf
8. Lee, A., Kawahara, T.: Recent development of open-source speech recognition engine Julius. In: Asia-Pacific Signal and Information Processing Association, pp. 131–137. Annual Summit and Conference, International Organizing Committee (2009)
9. Lewis, M., Gary, F., Charles, D.: Ethnologue: Languages of the World, 18th edn. Dallas, Texas (2015)
10. Onyenwe, I.: Developing Methods and Resources for Automated Processing of the African Language Igbo. Ph.D. thesis, University of Sheffield (2017)
11. Schultz, T., Waibel, A.: Language-independent and language-adaptive acoustic modeling for speech recognition. Speech Commun. **35**(1), 31–51 (2001)
12. Wells, J.: SAMPA computer readable phonetic alphabet. In: Handbook of Standards and Resources for Spoken Language Systems, vol. 4 (1997)
13. Young, S.J., Young, S.: The HTK hidden Markov model toolkit: design and philosophy. University of Cambridge, Department of Engineering (1993)