Chapter 8
A State-of-the-Art Review of Nigerian Languages Natural Language Processing Research

Toluwase Victor Asubiaro
https://orcid.org/0000-0003-0718-7739
University of Ibadan, Nigeria & Western University, Canada

Ebelechukwu Gloria Igwe
https://orcid.org/0000-0002-5180-5621
University of Ibadan, Nigeria

ABSTRACT

African languages, including those that are natives to Nigeria, are low-resource languages because they lack basic computing resources such as language-dependent hardware keyboard. Speakers of these low-resource languages are therefore unfairly deprived of information access on the internet. There is no information about the level of progress that has been made on the computation of Nigerian languages. Hence, this chapter presents a state-of-the-art review of Nigerian languages natural language processing. The review reveals that only four Nigerian languages; Hausa, Ibibio, Igbo, and Yoruba have been significantly studied in published NLP papers. Creating alternatives to hardware keyboard is one of the most popular research areas, and means such as automatic diacritics restoration, virtual keyboard, and optical character recognition have been explored. There was also an inclination towards speech and computational morphological analysis. Resource development and knowledge representation modeling of the languages using rapid resource development and cross-lingual methods are recommended.

INTRODUCTION

The inclusion of countries in the information society is importantly determined by their ability to access, create, and use information on the global information highway. Most prominent in the global report on measuring the information society is the annual report of the International Telecommunication Union

DOI: 10.4018/978-1-7998-3468-7.ch008

(ITU) which is pivoted on gadget and infrastructure-focused metrics such as internet use, telephone penetration, mobile telephone use, access to computer and other ICTs, broadband access, mobile signal availability, internet bandwidth size and internet traffic. Recent reports show that developing countries, which also belong to the *have-nots* in the digital divide, are improving on the ITU's information society metrics, though questions arise about the impact of the recorded progress on the developing countries' socio-economic development. Studies have suggested that the problem of inequalities in access to information have continued, even in the information era and despite the progress made by the developing countries as reported in the annual *Measuring the Information Society* reports of the ITU. Jansen and Sellar (2008) for instance, noted that, "… despite all the advances made in promoting access" through "… ICT and internet -the same familiar inequalities persist". Perhaps, the present metrics and efforts at bridging the digital divide do not include the most important type of access to information, which is in the mothers' language of the developing countries.

The importance of information access in the mothers' languages of the developing countries on bridging the digital divide has been expressed by earlier researchers using different terms and concepts. In explicit terms, Yu (2002), stated that "…barrier to digital participation is language". Adegbola (2017) described access to information in languages that are spoken by the local population of the developing countries as "the last six inches" of the digital divide bridge. Osborn (2010) recommended glocalization which is "the adaptation of digital information and contents to the local modes of communication, culture and standards", with much emphasis on provision of services and content creation in local languages (language access) as a panacea to bridging the digital divide. Borgman (2000) in "thinking locally, acting globally", suggested the development of customized or human-centered information systems that is dependent on age, expertise, language and other socio-demographic characteristics of individuals. These studies and others have recommended that language access to information is sacrosanct to bridging the digital divide.

Languages that are spoken by the countries in the *have-not* of the digital divide are regarded as resource-scarce languages. Resource-scarcity for languages in the digital age is used in tandem with other terms such as low-resource, resource-poor, under-resourced, resource-limited and resource-constrained to describe the dearth of computer resources such as large and accurate text and speech corpora, analytical tools (part-of-speech (POS) tagger, chunking systems, parsers, stemmers, lemmatizers syllabicators), inputting tools (keyboards, speech-to-text systems) and knowledge tools (models, machine translation (MT) models, computational grammar, morphology rules, etc) for the natural language processing (NLP) of such languages. NLP refers to the interdisciplinary field that draw knowledge from computer science, artificial intelligence, linguistics, statistics, and machine learning, and it focuses on analyzing and studying human languages (text and speech) with the aim of developing computer programs that can process human languages in human-like format. Availability of resources for a language, and subsequently the intensity of its NLP research, strongly correlates with the availability of digital application and contents for and in the language. Better still, languages in the *have* divide of the digital world have plenty of resources and relatively high number of NLP research than those in the *have-nots*. One of the gaps in literature is the review of NLP research of the Nigerian languages to evaluate the progress in bridging the digital language divide. This book chapter, therefore, provides a state-of-the-art review of the developments that have been made on the NLP of Nigerian languages by thematically analyzing the content of publications on the NLP of the languages.

Nigeria is the most populous African country with over 200 million population and 400 indigenous languages, though only four (Hausa/Fulani 29%, Yoruba 21%, Igbo 18% and Ijaw 10% (CIA, 2016))