Computing Perplexity Values for Under-resourced Languages using n-gram and Deep Learning Approaches

BAYANG SOULOUKNA Jules Paulin *paulinbayang@gmail.com*
Faculty of Science/ Department of Mathematics and Computer Science
Laboratoire de Recherche en Informatique (LARI)
The University of Maroua

DAYANG Paul *pdayang@univ-ndere.cm*
Faculty of Science/Department of Mathematics and Computer Science
Laboratoire de Recherche en Informatique (LARI)
The University of Ngaoundéré

KOLYANG *kolyang@univ-maroua.cm*
Higher Teachers' Training College/Department of Computer Science
Laboratoire de Recherche en Informatique (LARI)
The University of Maroua

WADOUFEY Abbel *wadouf@gmail.com*
Faculty of Science/Department of Mathematics and Computer Science
National Institute of Cartography, Cameroon
The University of Ngaoundéré

Abstract

The interactions between computers and human language, through the approach called natural languages processing, need a very good model describing the language and a large amount of data. But for under-resourcedlanguages, however due to lack of resources (texts resources), it becomes challenging to devise a good model adapted for minority languages. To cope with this issue, in this paper, we focus on the collection of data for the construction of a language model adapted to poorly endowed languages. Firstly, we describe the concept of under-resourced languages and difficulties related to the digital processing of those languages. To illustrate our model, we collect some text data of Tpuri an African language spoken in Cameroon and Chad. For the collection, we used diverse sources like existing printed documents. Our dataset contains 1640128 words and 108553 sentences. With the collected dataset, two main stemming approaches(n-gram and recurrent neural network)have been evaluated. The perplexity values have been computed in order to judge how good language model is according to the characteristics of under-resourced languages. For the statistical n-gram language model, we obtainedthe perplexity valueof 420.01 for bigram and 270.45 for trigram. Relying on a linear interpolation with λs= [0.2, 0.2, 0.4, 0.2], a best perplexity value of 56.74 could be determined. We also obtained a best perplexity equal to 47.21 with Laplace smoothing using 4-grams, when λ has a value of 0.03.Implementing a recurrent neural network model using the multilayer perceptron (long short-term memory), we obtain a perplexity value of 77.18 which is to be considered as a better result.

Keywords: Language Model, Tpuri, n-grams, Neural Network, Long Short-term Memory, Multilayer Perceptron, Perplexity.

1. INTRODUCTION

In natural language processing, data are the key success factors but they are sometimes complicated to acquire (Bellegarda & Monz, 2015). The language dataset made oftext corpus, lexicons, pronunciation dictionary etc. are used to build prediction models for the test and

validation of automatic word processing applications (Information retrieval, automatic summary, Automatic Speech Recognition System...)(Vu-minhet al., 2006.). These resources are therefore very important for the automatic natural language processing which relied on statistical approaches that often require a very large amount of data to process learning models(El-Haj et al., 2015).

In the world, we have more than 7000languages(Paolillo, 2006) and most of them are considered as highly endowed. It is very challenging to get appropriate language resources when we want to digitalise these languages.Africa and Asia combined have two-thirds of the world's SIL languages (over 4,300 languages). Africa with nearly 2000 languages spoken(Paolillo, 2006), remains the continent with the largest number of endangered languages and whose computerization has the greatest delay in acquiring the construction of technologies usable for these languages. This can be explained by the lack of resources in large quantities for these languages, the number of speakers which are less and less numerous and in-depth studies for these. Most African and Asian languages are considered like language π (for poorly endowed language) and language μ (for medium endowed)(Besacier et al., 2014). To protect these languages, it is becoming important to propose a set of tools that will contribute to the computerisation process of the so-called poorly endowed languages.

In order to safeguard languages, it becomes important to offer solutions and tools for these different languages. Several works in this direction have been proposed for poorly endowed languages(Le et al., 2003)(Nimaan et al., 2006)(De Wet et al., 2016). Butmuchremains to bedone. It is in this perspective that our work is presented, which will consist in evaluating and comparing the perplexity of language models constructed after the collection of the textual resources collected and pre-processed.

Perplexity can be definedas a measurement of how well a probability model predicts a sample; it can be explained by $\text{Perplexity} = 2^{H_m(W_1..W_n)}$ Where $H_m(W_1..W_n)$ is the Cross entropy given by$H_m(w_1..W_n) = \frac{-1}{n} log_2 P_m(W_1..W_n)$. Lower perplexity mean that the model language is good.

In this paper, we review the processing of under-resourced languages, some aspects which are used to classify languages, the challenges concerning these languages and some obstacles related to processing of these languages, particularly for African languages. Secondly, we present Tpuri language that we have chosen with its characteristics and explain how we make a collection of our text corpus and distribution. After this, we use our corpus to build two types of linguistic models, one of them concern n gram model, where we use bi gram, 3-gram, we use linear interpolationand n-gram with Laplace smoothing and evaluate the perplexity of our model. The second type of our linguistic model is a neural network-based model, particularly Multi-Layer perceptron and Long short-term Memory based model that we alsoevaluate perplexity of each linguistic model. Lastly, we comparedifferent perplexity results of our models built and present the prospects for computerization for the African languages in general.

2. PROCESSING OF UNDER-RESOURCED LANGUAGES

Usually, under-represented languages or under-resourced languages(El-Haj et al., 2015; Tomasz, 2018), poorly endowed languages(Bellegarda & Monz, 2015; Eshkol & Antoine, 2017; Esuli et al., 2016) or very little equipped languages are considered as minority languages. Importantly, these terms describe languages characterised by the lack of a unique writing system or stable orthography, limited presence on the web, lack of linguistic expertise and lack of electronic resources for speech and language. Nevertheless, there are languages having a high number of speakers but classified among the said very little equipped languages. For instance, it is the case of Hindi and the Khmer (Le et al., 2003). In our context, we consider as little equipped languages those with limited and reduced digital resources.

More so, the main reasons why we deal with language processing are (Gauthier et al., 2016) (Nimaan et al., 2006):

- Assistance in writing based mainly on a dictionary, orthographical and grammatical corrector;
- Computer-assisted language teaching;
- Optical character recognition;
- Treatment of oral examination, voice synthesis, machine translation, voice recognition and automatic transcription.

These various purposes resulted in classifying the languages according to the three index groups (Caelen et al., 2006):

- *Languages-π* which are known as little equipped languages, many African, Asian and South American languages;
- *Languages-μ* which are as fairly equipped languages Like Vietnamese, Amharic, Swahili…
- *Languages-τ* which are very well-equipped languages; In this category, one finds the languages strongly spoken like English, French, German, Japanese…

2.1. Resources for Poorly Endowed Languages in the Digital World

Typically, in order to collect resources to build language models for very well-endowed languages, an interesting approach will be to retrieve very large amounts of dataset from websites in the given language and filtering them so tomake them useful for the system to be constructed. These textual data can then be used on the one hand to process statistical models of the language, and on the other hand to obtain a corpus that can then be pronounced by speakers in order to build up a consistent signal base(Etman & Beex, 2015) . The data thus collected will be processed in particular for the conversion of encodings and also for segmentation (Peter Jackson, 2004). Unfortunately for poorly endowed languages, they do not have a strong representation on the web; one of the characteristics of poorly endowed languages is the absence of a fixed graphical design. These languages remain mostly oral; since the whole culture is largely based on orality.

African languages are mostly thoseclassifiedas languages-π. These languages encounter several challengeslinked in particular to the portability of the models for building automatic speech recognition and transcription. The main difficulties described in the literature are the alphabet, the morphosyntax, the phonology and the orthography.

The alphabetis one of the major difficulties of languages which are little represented in the digital world. Several languages have at the same time an alphabet based on Arabic and another based on Latin. Furthermore, some characters of the *phonetic alphabet International* (API) have been added. These characters allow the transcription of the sounds of many under-represented languages. For example, the Urdu language in India (the fourth most spoken language in the world) has special writing letterswith 58 characters, while the Amharic language in Ethiopia have more than 200 symbols to represent syllables (Pellegrini & Lamel, 2006). At the level of the textprocessing, the variety of character sets becomes a huge challengedespite the definedcharacter encodings offered by the ISO/IEC 8859 standards.The existing standards aim to improve and to define characters for representing languages but the African languages still remain the most absent (Camara et al., 2004). Therefore, several of these languages cannot be represented by the existing standards due to the absence of signs.

In the morphosyntax, we deal with word-formation (morphology) and sentence-formation (syntax). With the lack or the insufficiency of morphosyntax rules, it becomes complex to carry out pre-treatments on digital little equipped languages for the automatic language processing using the approaches of stemming and lemmatization. Some information in several languages remains non-existent or very ambiguous.

Phonology is the study of the patterns of sounds in a language and across languages. Basically, signs are important to study how languages systematically organise their sounds. Therefore, the

absence or the insufficiencyof textual corpora can hinder the efficient and systematic analysis of sounds in order to produce patterns for processing an automatic speech recognition system.

Orthography defines a framework of rules and conventions for writing a language. It is a specification of how words of a language are mapped to and from a particular script. The prerequisitesare the script or a set of well-defined characters. Unfortunately, many African languages remain oral in majority without a formal orthography. The lack of an orthography is a big barrier for processing language.

2.2. Collection Techniques for Creating Resources for African Languages

In general, there are three main approaches that we can use to increase the size of digital resources or to generate text corpus (Mahtout, 2014).These methods can be combined for a better collection.

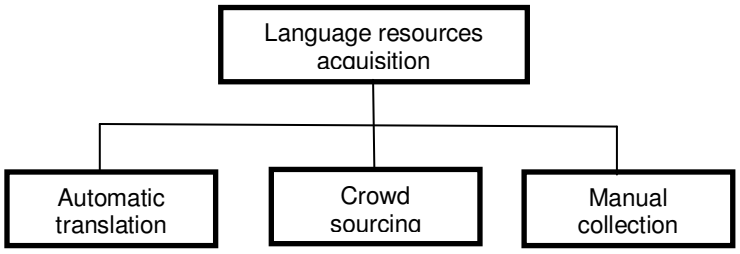

FIGURE 1: Techniques for Creating Language Resources.

- **Automatic translation** consists in translating corpora from a source language to a target language. Knowing that some languages have very large corpora, this approach can allow to obtain large quantity but with poor quality;

- **Crowdsourcing** is used to acquire resources very quickly but with also relatively poor quality (Le et al., 2003);

- **Manual collection** remains the most reliable and secure method of obtaining language resources of unquestionable quality, but unfortunately it requires too much human resources and time. In addition, language experts are needed and also the collection should ideally be carried out among native language speakers.

Some studies presented many approaches that were used to collect dataset for African languages. Some unannotated dataset for modelling the Swahili language was collected by combiningopen available data(Shikali & Mokhosi, 2020) and also data translated from English using Google Translate. These data were classified into three types; unannotated Swahili which are sentences, a list of syllables and four groups of related words. Nowadays, Swahili dataset can also be used for some natural language processing tasks like machine translation, sentiments analysis or part-of-speech tagging. Despite the fact Swahili is spoken by more than 100 million persons, it still belongsto low-resourced languages.

Concerning South African languages, a dataset for comparable evaluation of machine translation(McKellar & Puttkammer, 2020) is proposed, these dataset encompasses 11 South African languages. Each language has 500 sources made by professional human translators. Furthermore, some tools like language lemmatisers, part-of-speech taggers and morphological decomposers have been developed for10South African languages by annotating approximately 50000 tokens per language(Eiselen & Puttkammer, 2014).

For SATOS (Swahili, Amharic, Tigrigna, Oromo, Somali languages) machine translation project(Lakew et al., 2020) , some sources like Bible, JW300, Ted talks corpus and Wikipedia dumps were used to collect text data.

Another work done in the framework of collecting data is the annotated corpus of Guinean Maninka, called Maninka Reference Corpus(Vydrin et al., 2016)(Vydrin et al., 2014),which includes two main groups of corpus, the first one is written in Latin-based graphics (which more around 800,000 words, and the other one comprised texts in N'ko alphabet with more than 3,000,000 words. The texts were mainly collected from private individuals who are rather active on Internet and the Bible.

Several other works concerning the collection of resources forAfrican languages have been carried out in particular for Arabic (Brour & Benabbou, 2019; El-Haj et al., 2015), Igbo (Onyenwe, 2017), Mbochi (Rialland et al., n.d.), Amharic(App et al., 2016), Bambara (Tapo et al., 2014; Vydrin et al., 2016) nevertheless, much still remains to do for African languages.

3. DESCRIPTION OF THE TPURI LANGUAGE

Tpuri or Toupouri is an African language which is spoken in two neighbouring countries, Cameroon (in its far north region) and Chad (in its southernpart) by about 234,000 persons[1]. Like most of the languages of this region, Tpuri belongs to the Niger-Congo family and is written using the Latin alphabet. This language has four dialects: Podokge (Tui-pod), Bang-Were (Tui-bar), Faale-Piyew (Tui-faa) and Bang-Ling (Tui-ban). Referring to some existing studies (Ruelland, 1992), it can be considered as a poorly endowed language (Languages - π) due to the lack of sufficient resources. Some studies have been conducted concerning this language, in particular a Tupuri-French-English dictionary has been published (Ruelland, 1992; Taiwé, 2010), (Ruelland, 1998)as a book describing the language as well as a mini lexicon (Taiwé, 2010). This language remains highly oral and classified as a developing language according to l'ethnologue[2].

Classified as an under-represented language, Tpuri does not have enough resources. In addition to the insufficiency, the few existing text corpus does not necessarily use the same alphabet, the phonology and spelling are not harmonised. The main source of our corpus is derived from the Tpuri's Bible produced in 2006 by the BSC4. Other sources are the book *Ɗiggi, Siŋri wo* [19], a collection of poems and stories from a local radio station (radio daana) and some proverbs from social networks. The different sources are summarised inTable1.

Corpus	Number of words	Number of sentences	Authors
Tupuri Bible	1144187	8691	Alliance Biblique du Cameroun(ABC)
Jehovah's witness texts	11198	3527	www.jw.org/tui
Ɗiggi/Siŋri wo	28740	1241	Kolyang Dina Taiwe
Tales and proverbs	24181	9841	ProverbesTupuri
Naatogod Baa	431822	7083	EgliseFraternelleLutherienne
Total	**1640128**	**108553**	

TABLE 1 : Sources and amount of corpus.

[1]This evaluation was made since 1984, and is use by some authors

[2]https://www.ethnologue.com/language/tuiconsulted on December 22, 2020

4. LANGUAGE MODELLING OF TPURI

To process language automatically, there exist threemain groups of stemming algorithms: truncating methods, statistical methods and mixed methods. Figure 2 gives an overview of these algorithms(Jivani & others, 2011).

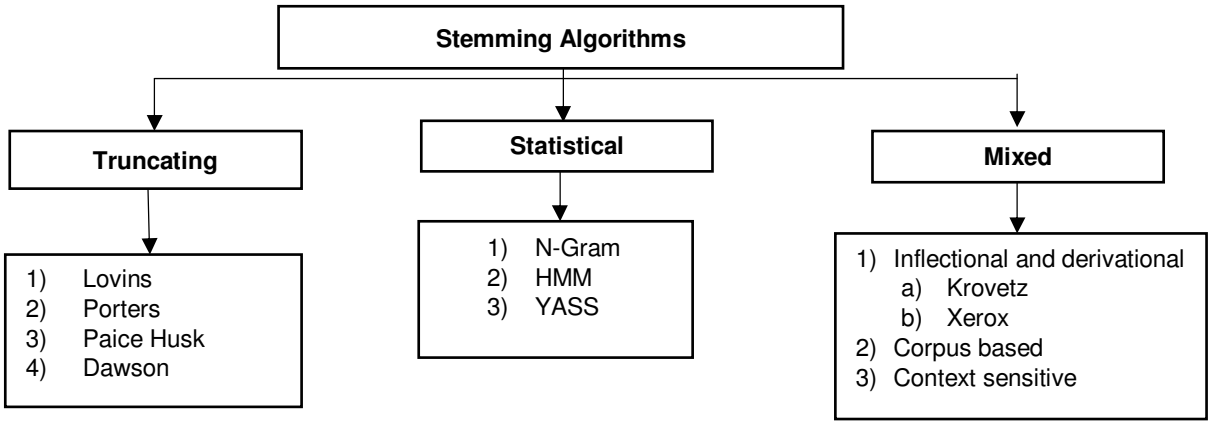

FIGURE 2: Overview of the existing stemming algorithms [20].

According to the literature, the best stemming approaches belong to the group of mixed algorithms (truncating and statistical). For our work, we have chosen thestatistical approachcalled*n-gram*and neural model. The model of the language that we build is essentially based on a corpus distributed as shown inTable 2.

The Tpuri corpus is subdivided in three parts. Two thirds were allocated for the training and the remainder subdivided in two parts; the training and the test corpus. Table 2 presents an exact repartition of dataset.

Corpus repartition	Corpus size (number of words)
Test Corpus	273360
Training Corpus	993414
Validation corpus	273354

TABLE 2: Training data, test data and validation data.

4.1 Text processing using n-gram Model

For this work, we proceed with some pre-processing in four steps, which are:

1. Removing punctuations and numbers;
2. Replacing multiple spaces with a single space;
3. Converting text to lowercase;
4. Tokenizing the text into individual words.

By carrying out this pre-processing above, we standardize the Tpuri corpus. In the language Tpuri, several words are made up of a single character. What amounts losing information if we remove the words having only one character.

4.1.1. N-gram models without Smoothing

Generally, in the n-gram model(Esuli et al., 2016), the n^{th} word is generated starting from the *n-1*position. In a formal way, it will be said that this model checks the assumption that the property of Markov of *n-1* order is true.

For a sequence of words $W = w_1, w_2 \dots w_n$, with $n = 1,2, \dots, n$ one will have $P(W_n|W_1^{n-1}) = P(W_n|W_{n-N+1}^{n-1})$

with *n* the number of grams and N is the count of all tokens in the test set.

The n-gram model without smoothing calculates the probabilities of the grams starting from the relative frequencies of the n-gram concerned.

In general, n-gram formula is definedas:

$$P(W_n|W_{n-N+1}^{n-1}) = \frac{C(W_{n-N+1}^{n-1}W_n)}{C(W_{n-N+1}^{n-1})}$$

With $C(W_{n-N+1}^{n-1}W_n)$the frequency of N gram$W_{n-N+1}^{n-1}W_n$, and $C(W_{n-N+1}^{n-1})$ the frequency of N-1 gram W_{n-N+1}^{n-1}.

Perplexity of unigram, bigram, trigram and quadrigram without smoothing of our corpus. The results obtained are presented in Table 3.

N-gram	Perplexity
1-gram model	294.02
2-gram model	212.13
3-gram model	177.53
4-gram model	**109.18**

TABLE 2: n-gram perplexity without smoothing.

Since Tpuri is a poorly endowed language, we are interested in improving the estimation of less frequent word probabilities using two smoothing: Laplace smoothing and interpolation smoothing.

4.1.2. N-gram models with Laplace Smoothing

We use Laplace smoothing for building our model language. If the corpus contains only one small part of the possible n-grams, the largest share of the mass of probability is distributed on the n-gram that is not observed. We vary the number of grams as well as the parameter lambda changing from 0.01 to 1. The table below presents the perplexities obtained.

We obtain the best perplexity with the value **47.21**, for the 4-grams when lambda equals to 0.01. It is noted that for N grams best perplexity is obtained when lambda is equal to 1. Beyond 4 grams, perplexity starts to be degraded gradually.

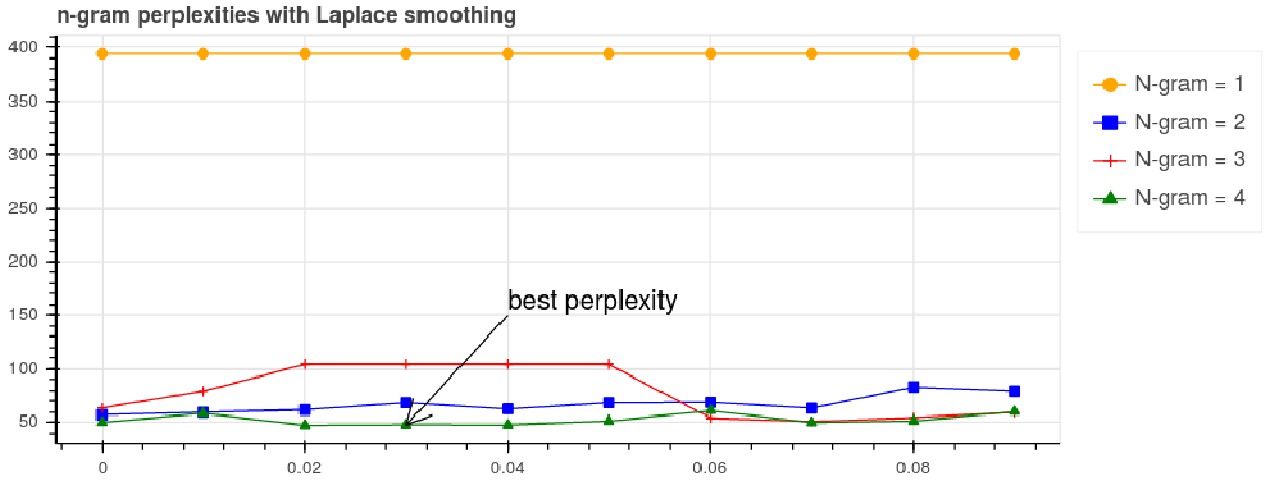

FIGURE 3: n-gram perplexities with Laplace smoothing.

4.1.3. Jelinek-Mercer (Chen & Rosenfeld, 2000)Smoothing

We build an interpolated language model using brute-force approach to find λs to maximize the probability(Chen & Goodman, 1999).The result obtained is presented in Table 5 with sum of λs equals to 1.

For our work, we use a simple interpolation formula which is given below:

$$P(W_n|W_{n-1}W_{n-2}W_{n-3}) = \lambda_1 P(W_n|W_{n-1}W_{n-2}W_{n-3}) + \lambda_2 P(W_n|W_{n-1}W_{n-2}) + \lambda_3 P(W_n|W_{n-1}) + \lambda_4 P(W_n)$$

We find λsthat maximize the probability using formula is given by:

$$logP(W_1..W_n|M(\lambda_1..\lambda_k)) = \lambda \log P_{M(\lambda_1,\dots\lambda_k)}(W_i|W_{i-1})$$

After setting the values of Lambda as followed, *unigram=0.2*, *bigram=0.2*, *trigram=0.3* and *quadrigram=0.3* we obtain a best perplexity value which is **56.7.**

Lambda set	Perplexities of corpus test	Lambda set	Perplexities of corpus test
[0.1, 0.1, 0.1,0.7]	138,96	[0.2, 0.2, 0.2, 0.4]	128,11
[0.2, 0.1, 0.1,0.6]	186,92	[0.2, 0.2, 0.3, 0.3]	120,49
[0.3, 0.1, 0.1, 0.5]	138,96	[0.2, 0.2, 0.4, 0.2]	56.74
[0.4, 0.1, 0.1, 0.4]	128,11	[0.2, 0.2, 0.5, 0.1]	110,81
[0.5, 0.1, 0.1, 0.3]	120,49	[0.2, 0.3, 0.4, 0.1]	57,81
[0.6, 0.1, 0.1, 0.2]	114,92	[0.2, 0.4, 0.3, 0.1]	82,92
[0.7, 0.1, 0.1, 0.1]	110,81	[0.2, 0.5, 0.2, 0.1]	99,24
[0.6, 0.2, 0.1, 0.1]	107,83	[0.2, 0.6, 0.1, 0.1]	118,88
[0.5, 0.3, 0.1, 0.1]	105,77	[0.3, 0.5, 0.1, 0.1]	61,42
[0.4, 0.4, 0.1, 0.1]	104,53	[0.4, 0.4, 0.1, 0.1]	68,83
[0.3, 0.5, 0.1, 0.1]	104,06	[0.5, 0.3, 0.1, 0.1]	74,85
[0.2, 0.6, 0.1, 0.1]	104,38	[0.6, 0.1, 0.2, 0.1]	56,74
[0.1, 0.7, 0.1, 0.1]	105,55	[0.4, 0.1, 0.4, 0.1]	58,68
[0.1, 0.6, 0.2, 0.1]	107,74	[0.4, 0.3, 0.2, 0.1]	59,82

[0.1, 0.5, 0.3, 0.1]	111,27	[0.4, 0.2, 0.3, 0.1]	65,42
[0.1, 0.4, 0.4, 0.1]	116,72	[0.5, 0.1, 0.3, 0.1]	59,96
[0.1, 0.3, 0.5, 0.1]	125,44	[0.1, 0.1, 0.4, 0.4]	57,75
[0.1, 0.2, 0.6, 0.1]	141,36	[0.1, 0.1, 0.5, 0.3]	83,83
[0.1, 0.1, 0.7, 0.1]	138,96	[0.1, 0.1, 0.6, 0.2]	88,76

TABLE 3: Perplexities for the dataset validation.

We obtain better perplexities for our models with smoothing compared to the model without smoothing particularly that of the smoothing of Laplace when *n=4*. These smoothing's techniques made it possible to allot to words which were less frequent in the corpus to have a probability of being also observed.

4.2. Text Processing using Neural Network Models

There exist diverse approaches to implement neural models. In this case, we have chosen the Multilayer Perceptron (MLP) which belongs to feedforward artificial neural network and the Long Short-Term Memory (LSTM) which is an artificial recurrent neural network.

4.2.1. Multilayer Perceptron

The architecture we have used for the multi-layerperceptron describedin (Bengio et al., 2003). Batch size used for this training and evaluation model is 256; with 20 epochs and 200 hidden units;the order of model is 5, which means a 4-word sequence followed by a 1-word prediction.The perplexity values obtained for the text distribution are:

- final training perplexity: 148.06;
- final validation perplexity: 140.30;
- final test perplexity: 125.12.

4.2.2. Long Short-Term Memory

For evaluating language models, the LSTM approach proposed in literatures [22], [23] [24] seem to be the most suitable. It is a recurrent neural network which is able to access and retain relevant information over a long interval. The vanish gradient problem is reduced through the introduction of multiplicative logic gates. We proposed a model based on the perceptron multilayer. On the one hand, we used the LSTM to deal with the performance issue during the processing of natural languages.

For the training model, we varied some parameters as followed:

- Number of hidden neurons: 100- 1500 (step size: 100, 300, etc.),
- Embedding size: 100 - 500 (step size: 210, 220, etc.),
- Decay rate: 0.5 - 0.9 (step size: 0.51, 0.52, etc.),
- Number of epochs before starting to decay the learning rate: 5 - 20 (step size: 5, 6, etc.).

After finishing the first round of the random search, the model with the best validation perplexity was chosen for further testing. In the second round of random search, we varied the following parameters: *embedding dropout probability*, *LSTM input*, *recurrent* and *output dropout probabilities*, *gradient clipping norm*. At the end of the process, we obtained the best results for the validation perplexity whichequals to 62.12 and test perplexitywhich equalsto 77.18 with 1200 hidden neurons and an embedding size of 360. The results are summarised in Table 6.

Models	Training perplexity	Validation perplexity	Test Perplexity
MLP	148.06	140.30	125.12
LSTM	57.20	62.12	77.18

TABLE 4: Summary of perplexities.

LSTM Model presents of better results without doubt due to the fact that it is betteradapted for the text than PML.Perplexity of model LSTM less good than models described in (Lau et al., 2017) and(De Mattei et al., 2020) . It can be explained by the fact that these models use great quantities of data compared to the Tpuri corpus and also quality of the data used.In comparison with (Bengio et al., 2003) the training and validation perplexities are lower than the paper presented and they can go even lower with more training, but test perplexities are a little bit higher.

N gram models with smoothing present better perplexities compared to models containing networks of neurons. That can be due to quality and also quantity of data. We have to make here with a little equipped language of which the first challenge remainder is dataset collection.

5. CONCLUSION

In this paper, we obtained a best result with n grams' models using Laplace smoothing. The lack of digital resources makes it very challenging to process African languages, those known asunder-resourced languages.After characterising the so-called under-resourced languages and identifying the challenges to deal with, we computed perplexity values using both the n-gram and deep learning approaches. Our aim was to generate perplexity values of three main groups of text data (training, test and validation). As final results to be considered, n-gram models using Laplace smoothing with 4 grams and lambda equal to 0.01 have shown better results over othersn-gram based models and neural network models for under-represented languages. The result of neural model can be explained by the and quantity of available data. For further work, we want to enlarge the text corpus and propose some core technologies for treatment like lemmatisers, POS tagger, morphological decomposer for the Tpuri language with the final aim of building an automatic speech recognition system that will help to gather more accurate data.

6. REFERENCES

App, L. M. D., Blachon, D., Gauthier, E., & Besacier, L. (2016). *Parallel Speech Collection for Under-resourced Language Studies Using the Parallel Speech Collection for Under-resourced Language Studies using the L ig -A ikuma Mobile Device App*. *December*. https://doi.org/10.1016/j.procs.2016.04.030

Bellegarda, J. R., & Monz, C. (2015). State of the art in statistical methods for language and speech processing. *Computer Speech & Language*, *35*, 163–184. https://doi.org/10.1016/j.csl.2015.07.001

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *The Journal of Machine Learning Research*, *3*, 1137–1155.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*(1), 85–100. https://doi.org/10.1016/j.specom.2013.07.008

Brour, M., & Benabbou, A. (2019). ATLASLang NMT : Arabic text language into Arabic sign language neural machine translation. *Journal of King Saud University - Computer and Information Sciences*, *xxxx*. https://doi.org/10.1016/j.jksuci.2019.07.006

Caelen, J., Besacier, L., Bigi, B., Boitet, M. C., Mori, M. R. De, Haton, M. J., Berment, M. V., Caelen, M. J., & Besacier, M. L. (2006). *Reconnaissance automatique de la parole pour des langues peu dotées*.

Camara, É., Ndamba, J., Nstadi, C., Rey, V., & Véronis, J. (2004). Traitement informatique des langues africaines. *Documents ALAF-ALAI, Paris, CNRS*.

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, *13*(4), 359–394.

Chen, S. F., & Rosenfeld, R. (2000). A survey of smoothing techniques for ME models. *IEEE Transactions on Speech and Audio Processing*, *8*(1), 37–50.

De Mattei, L., Cafagna, M., Dell'Orletta, F., Nissim, M., & Guerini, M. (2020). Geppetto carves italian into a language model. *ArXiv Preprint ArXiv:2004.14253*.

De Wet, F., Badenhorst, J., & Modipa, T. (2016). Developing Speech Resources from Parliamentary Data for South African English. *Procedia Computer Science*, *81*. https://doi.org/10.1016/j.procs.2016.04.028

Eiselen, R., & Puttkammer, M. J. (2014). Developing Text Resources for Ten South African Languages. *LREC*, 3698–3703.

El-Haj, M., Kruschwitz, U., & Fox, C. (2015). Creating language resources for under-resourced languages: methodologies, and experiments with Arabic. *Language Resources and Evaluation*, *49*(3), 549–580. https://doi.org/10.1007/s10579-014-9274-3

Eshkol, I., & Antoine, J.-Y. (2017). *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN) Actes de TALN 2017, volume 2 : articles courts*. *2*. http://taln2017.cnrs.fr/wp-content/uploads/2017/06/actes_TALN_2017-vol2.pdf#page=177

Esuli, A., Fagni, T., Fern, A. M., & National, I. (2016). *JaTeCS , a Java library focused on automatic text categorization*. 1–5.

Etman, A., & Beex, A. A. L. (2015). Language and Dialect Identification: A survey. *IntelliSys 2015 - Proceedings of 2015 SAI Intelligent Systems Conference*, *December*, 220–231. https://doi.org/10.1109/IntelliSys.2015.7361147

Gauthier, E., Besacier, L., & Voisin, S. (2016). Automatic Speech Recognition for African Languages with Vowel Length Contrast. *Procedia Computer Science*, *81*, 136–143. https://doi.org/10.1016/j.procs.2016.04.041

Jivani, A. G., & others. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, *2*(6), 1930–1938.

Lakew, S. M., Negri, M., & Turchi, M. (2020). *L OW -R ESOURCE N EURAL M ACHINE T RANSLATION :* 1–10.

Lau, J. H., Baldwin, T., & Cohn, T. (2017). Topically driven neural language model. *ArXiv Preprint ArXiv:1704.08012*.

Le, V. B., Bigi, B., Besacier, L., & Castelli, E. (2003). Using the Web for fast language model construction in minority languages. *EUROSPEECH 2003 - 8th European Conference on Speech Communication and Technology*, 3117–3120.

Mahtout, M. (2014). A Methodology for semi-automatic structuring of a bilingual lexicographical corpus: the French-Kabyle case (Méthodologie pour la structuration semi-automatique du corpus dans une perspective de traitement automatique des langues: le cas du dictionnaire fr. *TALN-RECITAL 2014 Workshop TALAf 2014: Traitement Automatique Des Langues Africaines (TALAf 2014: African Language Processing)*, 123–133.

McKellar, C. A., & Puttkammer, M. J. (2020). Dataset for comparable evaluation of machine translation between 11 South African languages. *Data in Brief*, *29*, 105146. https://doi.org/https://doi.org/10.1016/j.dib.2020.105146

Nimaan, A., Nocera, P., & Torres-Moreno, J.-M. (2006). Boîte à outils TAL pour des langues peu informatisées : le cas du somali. *Jadt*. http://lexicometrica.univ-paris3.fr/jadt/jadt2006/PDF/II-062.pdf

Onyenwe, I. E. (2017). *Developing methods and resources for automated processing of the african language igbo*. University of Sheffield.

Paolillo, J. C. (2006). Evaluating Language Statistics : The Ethnologue and Beyond A report prepared for the UNESCO Institute for Statistics. *Language*.

Pellegrini, T., & Lamel, L. (2006). Investigating automatic decomposition for ASR in less represented languages. *Ninth International Conference on Spoken Language Processing*.

Peter Jackson, Ni. M. (2004). Review of "Natural language processing for online applications: Text retrieval, extraction and categorization."*TerminologyTerminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *10*(1), 177–179. https://doi.org/10.1075/term.10.1.12dro

Rialland, A., Aborobongui, M. E., Adda-Decker, M., & Lamel, L. (n.d.). Mbochi: corpus oral, traitement automatique et exploration phonologique. *Jep-Taln-Recital 2012*, *1*, 1. http://anthology.aclweb.org/W/W12/W12-1301.pdf%5Cnhttp://aclweb.org/anthology//W/W12/W12-1301.pdf

Ruelland, S. (1992). *Description du parler tupuri de Mindaore (Tchad): phonologie, morphologie, syntaxe*.

Ruelland, S. (1998). *Dictionnaire Tupuri - Français - Anglais*. Peeters.

Shikali, C. S., & Mokhosi, R. (2020). Enhancing African low-resource languages: Swahili data for language modelling. *Data in Brief*, *31*, 105951. https://doi.org/https://doi.org/10.1016/j.dib.2020.105951

Taiwé, K. D. (2010). *Parlons Tpuri*. L'Harmattan.

Tapo, A. A., Coulibaly, B., Diarra, S., Homan, C., Kreutzer, J., Luger, S., Nagashima, A., Zampieri, M., & Leventhal, M. (2014). *Languages : A Case Study on Bambara*.

Tomasz. (2018). *Spoken Language Identification*. *July 2013*. https://doi.org/10.13140/RG.2.2.29465.62561

Vu-minh, Q., Besacier, L., Blanchon, H., & Bigi, B. (n.d.). *Modèle de langage sémantique pour la reconnaissance automatique de parole dans un contexte de traduction Mots clés-Key words 1 Introduction*.

Vydrin, V., Rovenchak, A., & Maslinsky, K. (2016). Maninka Reference Corpus: A Presentation. *TALAf 2016 : Traitement Automatique Des Langues Africaines (Écrit et Parole). Atelier JEP-TALN-RECITAL 2016 - Paris Le*. https://halshs.archives-ouvertes.fr/halshs-01358144

Vydrin, V., Umr-, C., Bp, M., & Cedex, V. (2014). *Projet des corpus écrits des langues manding : le bambara, le maninka 1*.