Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2803–2811
Marseille, 11–16 May 2020

SENCORPUS: A French-Wolof Parallel Corpus

Elhadji Mamadou Nguer, Alla Lo, Cheikh Bamba Dione, Sileye Oumar Ba, Moussa Lo
UVS, UGB, UiB, Dailymotion, UVS
Senegal, Senegal, Norway, France, Senegal
{elhadjimamadou.nguer,moussa.lo}@uvs.edu.sn, lo.alla@ugb.edu.sn, sileye.ba@dailymotion.com, dione.bamba@uib.no

Abstract

In this paper, we report efforts towards the acquisition and construction of a bilingual parallel corpus between French and Wolof, a Niger-Congo language belonging to the Northern branch of the Atlantic group. The corpus is constructed as part of the SYSNET3LOc project. It currently contains about 70,000 French-Wolof parallel sentences drawn on various sources from different domains. The paper discusses the data collection procedure, conversion, and alignment of the corpus as well as it's application as training data for neural machine translation. In fact, using this corpus, we were able to create word embedding models for Wolof with relatively good results. Currently, the corpus is being used to develop a neural machine translation model to translate French sentences into Wolof.
Keywords: Parallel corpus, low-resource language, neural machine translation, Wolof, word embeddings.

1. Introduction

Parallel corpora are valuable resources for bilingual lexicography and natural language processing (NLP) applications such as statistical (Brown et al., 1990) and neural (Sutskever et al., 2014a) machine translation (NMT). Using machine learning techniques, computer models can learn from parallel corpus data to translate texts with relatively good quality, as illustrated by Google Translate. For most of the resource-rich languages, there are already many parallel corpora such as Europarl (Koehn, 2005), the Bible translations collected and annotated by Resnik et al. (1999) and the OPUS corpus (Tiedemann, 2012). However, for low-resource languages such as Wolof and Fula, such corpora are virtually nonexistent and, to a certain extent, this limits NLP research on these languages. In fact, Google Translate does not currently provide modules for these languages. For Wolof, an LFG-based computational grammar (Dione, 2014) and a Universal Dependency treebank (Dione, 2019) have been recently developed. Besides, language resources and tools are scarce for that language.

To promote the development of low-resource African languages, the SYSNET3LOC project has set up a team, which comprises researchers in NLP and Artificial Intelligence from the Virtual University of Senegal (UVS), Gaston Berger University (UGB), Dailymotion, and the University of Bergen (UiB). The project aims to effectively use machine learning methods to implement automatic translation systems between local languages in Senegal and Western languages. To the speakers of these languages, this will open the door to great levels of knowledge about the world which they may not otherwise have access.

Achievement of this objective requires the setting-up of multilingual corpora of about several thousand or million parallel sentences between local languages in Senegal and Western languages. However, building such linguistic resources presents a number of challenges. First, parallel texts between Senegalese languages and Western languages are extremely scarce, and, even when they exist, they are often subject to copyright and licensing restrictions. Second, the size of such eventual resources are often relatively small, and this places serious constraints on training NLP applications such as neural machine translation. Third, the quality of the translations and representativeness of such corpora are of high relevance, meaning that as far as possible various domains and text genres may need to be included in such corpora, while a balanced mix of texts is typically hard to achieve. Finally, the creation of such corpora involves non-trivial tasks in terms of preprocessing the raw texts, removing noise and aligning the source text with it's translation(s).

The main local Senegalese languages we plan to include in SenCorpus are: Wolof, Fula and Bambara. As far the Western languages are concerned, our aim is to include French and English. It is important to note that, at the current state, SenCorpus mainly contains resources for French and Wolof.[1] Accordingly, in this paper, we will only discuss the French - Wolof parallel corpus. Nevertheless, the model described here is still applicable to the other languages that we seek to promote.

The remainder of the paper is organized as follows. Section 2. gives an overview of the data collection process, including preprocessing, data organization and translation of the raw texts. It also describes the conceptual model adopted to manage the corpus through a shared online platform. Section 3. discusses the general process of aligning the bilingual sentences obtained from the previous step. Section 4. provides some corpus statistical analyses. In section 5., two case studies illustrating the application of the corpus are outlined. First, we examine it's application to develop word embedding models for Wolof. Then, we analyse the use of the corpus to develop an LSTM-based model to translate between French and Wolof. Finally, conclusions are drawn in section 6..

2. Data Collection Process

The main modules of the corpus annotation procedures are shown in Figure 1.

The overall construction process is divided into 5 major steps. First, we identified the appropriate sources for the corpus data, which came in various formats (e.g. pdf, text,

Authors thank CEA MITIC/UGB for supporting this work.

[1] We have few resources available for Fula.

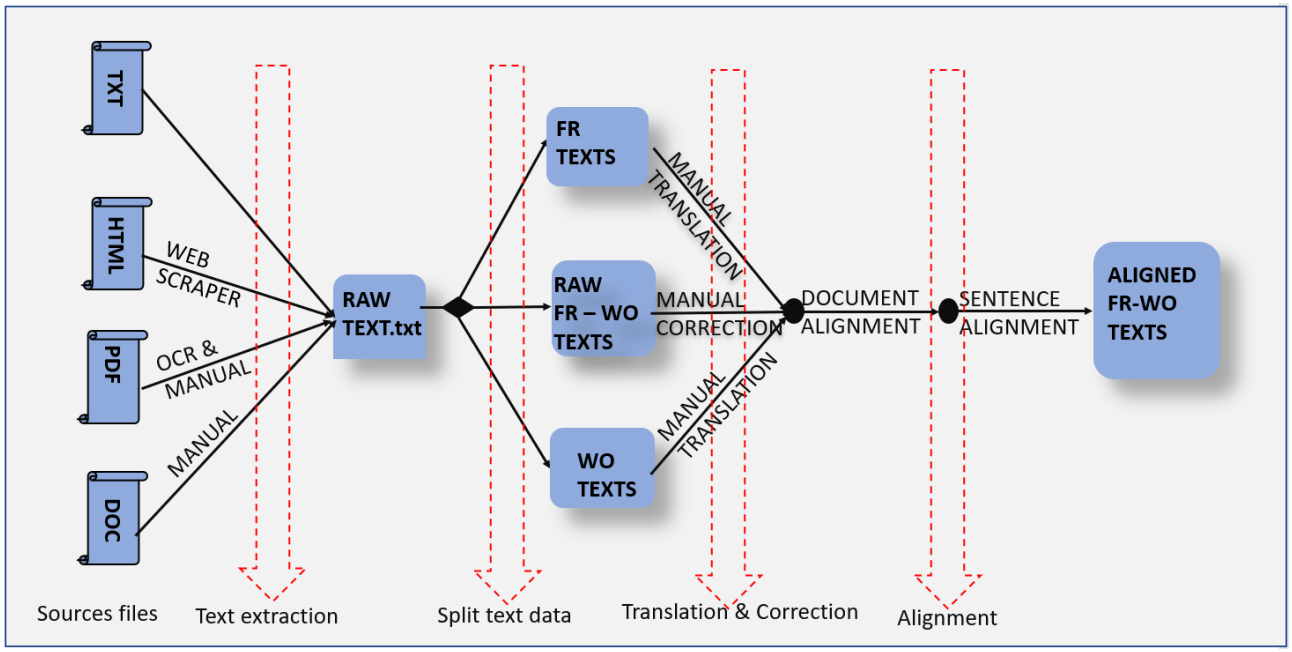


Figure 1: Data collection pipeline.

html and doc). Then, we proceeded to the content extraction. Next, we split the raw text data into both monolingual and bilingual texts. Subsequent to this, the monolingual texts were translated (e.g. French texts into Wolof). Likewise, the bilingual texts were manually corrected. The output of this stage consisted of bilingual (or multilingual) parallel documents, which were then aligned at the sentence level.

2.1. Corpus Data

During data collection, we placed special emphasis on the quality of the content and translation as well as the representativeness of the corpus. To ensure a good quality of the corpus, the sources are carefully selected and their translations manually verified. Also, special attention was paid to the intended applications. Even though our primary focus is on NLP applications, the corpus has to be designed in a way to also satisfy the need for human users (translators, linguists, teachers and students of foreign language, etc.). As for representativeness, our goal is to include texts from special domains (e.g. education, laws, religion, society, legend) as well as more general domains like agriculture, arts, cultures, history, geography, health, science. Regarding genres, we chose both fictional and non-fictional texts. As a general principle, we envisage to have for the final version a well distribution of the corpus sentences across the aforementioned domains.

2.2. Preprocessing

The first phase of the data collection process consisted in cleaning up the original material that we received from the different sources. Accordingly, we converted the various formats (e.g. HTML, RTF, DOC, and PDF) to plain text or JSON files. Original pdf files were first scanned, then proofread (where necessary) and corrected to ensure that the plain text document is complete and correct. For websites, we used crawling methods designed in form of Python scripts to extract bilingual documents. In addition, the texts were encoded according to international standards (UTF-8).

2.3. Data Organization and Translation

To facilitate the collaboration among actors involved in this project work, the corpus construction was carried out on the cloud following the steps outlined below.

The first step consisted in gathering the data in different folders, each reflecting a particular topic domain (e.g. sciences, religion, art, etc.). With such a structure, assigning a text document related to a specific domain was quite straightforward. Next, each collected resource was described in a table format which specified the domain of its content, the format (e.g. PDF, DOCX), the translation status (whether the translation was correct or not), the name of the translator, etc. For instance, as Figure 2 shows, the Bible and the Quran are entirely translated from French to Wolof. The short story "01-koumba" (Kesteloot and Dieng, 1989) is not translated into Wolof yet, while "02-àddina" has been translated, but the translation needs some revision. Vice versa, the French translation of *Doomi Golo: Nettali* (Diop, 2003) is not available yet.[2]

Furthermore, the documents were classified according to the language they are written in as well as their translation status. We then translated those documents that did not have a translation yet, and revised the other documents that have already been translated. Each bilingual document was converted in text format, corrected and validated.

[2]Recently, an English translation of this book has been made available by Diop et al. (2016).

N°	Title of the raw document	Content domain	State of the translation.	Folder	Source language	Target language
1	Le Coran	Religion	Fully translated	Rep1-Fr_Wo	French	Wolof
2	La bible	Religion	Fully translated	Rep1-Fr_Wo	French	Wolof
3	Dictionnaire Cissé et Al original	Generalit	Fully translated	Rep1-Fr_Wo	Wolof	French
5	Dictionnaire DILAF	Generalit	Fully translated	Rep1-Fr_Wo	Wolof	French
9	01-koumba	Legend	Not translated	Rep1-Fr	French	Wolof
10	02-àddina	Legend	Translation to revi	Rep1-Fr	French	Wolof
31	Brochure de terminologies juridiques r	Justice	Fully translated	Rep1-Fr_Wo	French	Wolof
32	Charte des droits de l'homme	Society	Fully translated	Rep1-Fr_Wo	French	Wolof

Figure 2: Examples of collected documents.

2.4. Corpus Management through a Shared Online Platform

SenCorpus is a multilingual corpus of parallel sentences. It is composed of a set of pairs (srcSent, targSent). srcSent is the source sentence and targSent the target sentence. The latter represents the translation of srcSent from the source language (srcLang) to the target language (targLang). At the current stage, most of the documents come from external sources that were already translated such as the religious texts mentioned above. In addition, some parallel sentences were extracted from dictionaries (Cissé, 1998; Mangeot and Enguehard, 2013), short stories (Lilyan and Cherif, 1983), booklets (internationale de la Francophonie, 2014), etc.

For a few documents (ca. 12,000 sentences),[3] however, translation from the source language (e.g. French or English) into a local Senegalese language (e.g. Wolof, Fula) was necessary. For the French-Wolof parallel corpus, translation was done by Wolof native speakers on a sentence-by-sentence basis. To ensure the quality of the translation, we set up several conditions as guidelines for individual translators. Each source sentence had to be translated into a single sentence in the given target language. The translation should not be in a word-by-word form and had to be faithful to the source text. It was not necessary for a source and a target sentence to have the same word length. The translation should preserve the meaning without adding complementary explanations (e.g. without trying to explain the meaning of the words or sentences).

Additional principles adopted during the corpus design included among other things that the corpus should be accessible from a shared online platform. Moreover, it had to be editable, collaboratively, by several users and exportable to various formats (Word, Excel, XML, text, Python, etc.).

The conceptual model defines the procedures through which a translator goes to complete all his tasks. The components of the model include users, languages, pivots or concepts, types of tasks and task elements. In this architecture, every user has a personal login, with an adaptable profile in accordance with his task. The personal login keeps track of information about the user such as her id, name, email, whether it is an active user or not, etc. Each user can be assigned one or more tasks (each having an explicit deadline and a completion date). A task is defined as a list of sentences to translate. In turn, a task element is a single sentence to translate. Each sentence is related to a pivot, which represents a specific concept. The use of the pivot is crucial in the sense that many sentences in different languages may refer to the same concept. Thus, sentences referring to the same concept will have the same pivot, i.e. use one and the same concept as a pivot. A new table of sentences is created for each newly added language.

[3]The 12,000 sentences were selected from the French-English parallel corpus available at https://www.manythings.org/anki/.

3. Automatic Alignment

One of the most challenging tasks for creating a parallel corpus is sentence alignment. Basically, this consists in extracting from a parallel corpus pair of sentences that are translations of one another. The task is not trivial for many reasons. For instance, some sections may be missing on one or the other side. Likewise, corresponding sections may not be translations of each other. Furthermore, a single sentence in one language may be translated as two or more sentences in another language. This means that operations like sentence deletions and substitutions may be required in order to reduce noise in the corpus.

Automatic sentence alignment methods generally fit into two classes: length-based vs. lexical-based. The former use length information to estimate alignment probabilities. A very popular length-based approach has been suggested by Gale and Church (1993). The intuition behind that approach is that "longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences" (Gale and Church, 1993, p. 75). Accordingly, a probabilistic score is assigned to proposed sentence pairs based on a distance measure. The most likely candidate of those pairs is selected using the maximum likelihood algorithm.

In contrast, lexical-based methods make use of lexical information to perform sentence alignment. Such information can be in the form of associative measures that are derived from bilingual dictionaries or translated parallel sentences. Lexical-based methods may also use internal lexical evidence, such as rare words, cognates, or word co-occurrences. An example of an alignment algorithm that is lexical-based is K-vec (Fung and Church, 1994).

In this work, we used a semi-automatic method to align Wolof and French texts at the sentence level. The next step after extracting the texts from PDF and other formats consisted in splitting the paragraphs of texts into sentences. This could be achieved by using a Perl script-based sentence-splitter. The script identified sentence boundaries based on characters like periods or question marks, capitalization and a list of exceptions. For Wolof, we supplied a customized version of the exception list provided for English. This was in order to account for cases that involve ambiguous sentence-ending markers (e.g. abbreviations).

For few documents like the Bible translations and the Quran, sentence alignment was quite straightforward, as these are already well-structured in terms of chapter and verse numbering. For instance, alignment of the French and Wolof versions of the Quran was done as follows. First, the surahs along with their verses were aligned and converted in JSON format. Then, another script took the JSON file, navigated through it surah by surah, verse by verse to extract the individual sentences contained in each verse.

Besides these religious texts, most of the other documents, however, came in unstructured format, which made it almost impossible to find correspondences. Because of the size of the corpus, it would be impractical to try to obtain a complete set of alignments by hand. Thus, we had to find some automated solutions for sentence alignment.

To assess the potential for an automated solution, we conducted an evaluation of three widely used open-source tools: hunalign (Varga et al., 2005), yasa (Lamraoui and Langlais, 2013) and champollion (Ma, 2006). hunalign is a hybrid algorithm that combines the dictionary and length-based methods. In contrast, yasa and champollion use lexical-based approaches. Evaluation of the three tools was done against a small set of manually aligned data from the corpus. We applied the methodology based on sentence-level alignment (every pair of aligned sentences does count), as proposed in Langlais et al. (1998). The evaluation results indicated that hunalign had satisfactory performance and was therefore used to semi-automatically align a part of the corpus (ca. 19k sentences). Besides, the remaining sentences of the corpus (except the religious texts and the part just described) were aligned manually. This is mainly because there was a substantial number of alignment errors produced by the alignment tools. A major cause of these errors was due to the noise contained in the texts. For instance, there were missing sections, texts without a corresponding translation or whose translation was not quite faithful.

The alignment tools hunalign and yasa are language-independent. They produce confidence scores, but these are not always reliable. This is true for both tools, but especially for yasa. Sentence pairs that clearly are unrelated sometimes get high scores if they happen to fill a gap between aligned sentences and have comparable length. In contrast, champollion does not give confidence scores.

The performance of hunalign can in principle be improved by supplying a dictionary of bilingual word or phrase pairs. We also tested this for hunalign by extracting a word list from the Wolof-French dictionary, but this had no positive effects on performance. In contrast, the performance seemed to drop. One potential reason for this is that the derived word list is too small to be potentially useful.

In a subsequent step, we manually revised all the automatically produced alignments, including those alignments obtained by running the scripts (i.e. the Bible and Quoran alignments) as well as those obtained from the alignment tools. We had to apply manual correction to ensure good alignment quality.

4. Analysis

4.1. Distribution of Domain Data

At the current stage, the French-Wolof parallel corpus is categorized in six major domains: education, general, laws, legend, religion and society. In the following, we provide a brief description and some statistics related to these domains.

- Education: The texts in this domain are acquired from teaching materials, such as language teaching resources and dictionaries. Totally, about 5200 parallel sentences were collected.
- General: This domain covers topics in various areas such as agriculture, history, geography, medicine, economy. It has around 25740 sentences.
- Laws: This domain consists of legal texts and has ca. 569 sentences.
- Legend: This domain contains narrative texts in form of short stories. The total number of parallel sentences is around 2162.
- Religion: This domain includes texts that are related to a religious tradition such as the Bible versions (Old Testament, New Testament, and the Jehovah's Witnesses texts), and the Quran. The religious texts contain about 35397 sentences.
- Society: This domains contains documents discussing various social issues such as human and civil rights, population, migration and immigration, development. Currently, there are 1966 parallel sentences assigned to this domain.

Figure 3 shows the distribution diagram of different domain data. From Table 1 and Figure 3, we can see that currently the domain *Religion* contains the most sentences (about 50% of the entire corpus). It is followed by *General*, which covers more than 35% of the corpus. The smallest portion is the parallel texts from *Society* and *Laws*. This is due to the fact that there are currently rare available resources in those domains.

4.2. Distribution of Topics

We were also interested in modeling the distribution of topics in our corpus. For this purpose, we used topic models,

Domains	Languages	Tokens	Average Length	Vocabulary	Sentences
Education	French	36467	7.0129	7603	5200
	Wolof	27869	5.3594	6597	
Religion	French	831972	23.5040	49764	35397
	Wolof	739375	20.8881	44301	
General	French	169666	6.5915	15925	25740
	Wolof	161921	6.2906	10731	
Laws	French	10016	17.6028	2738	569
	Wolof	9951	17.4886	2461	
Legend	French	27780	12.8492	6460	2162
	Wolof	26051	12.0495	5292	
Society	French	25391	12.9151	6398	1966
	Wolof	26266	13.3601	5412	

Table 1: Statistic summary of the French-Wolof parallel corpus.

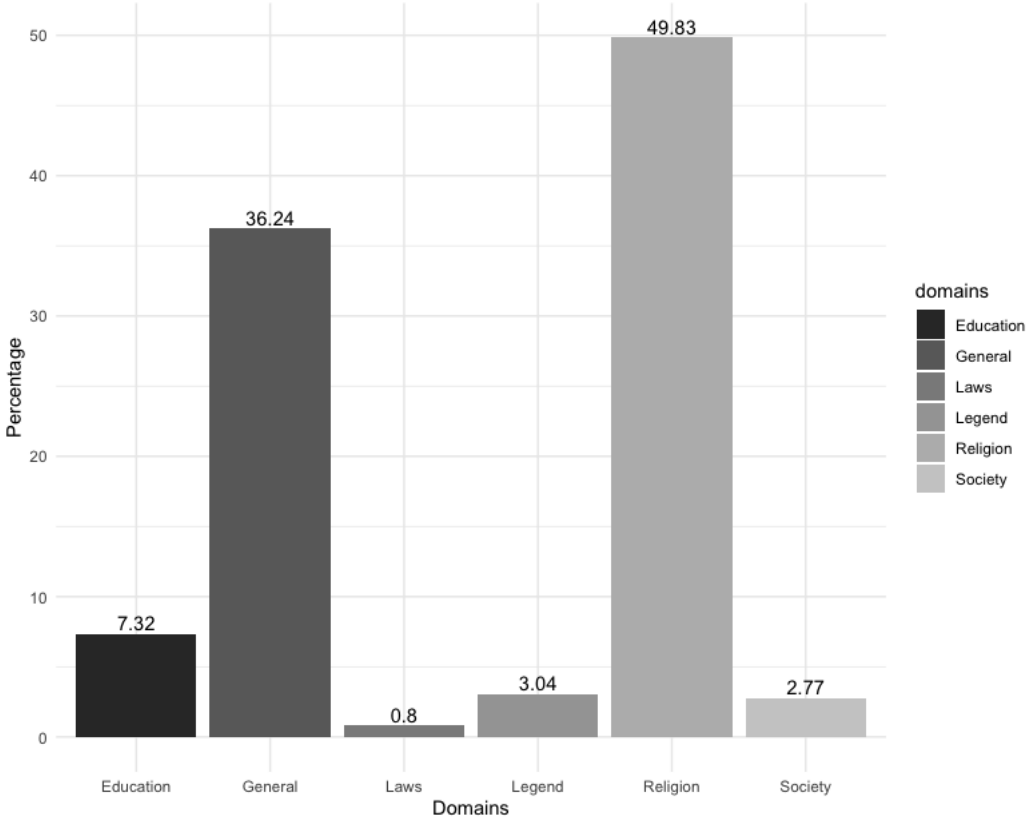


Figure 3: Distribution of domain data in the corpus.

which provide a relatively simple way to automatically discover and extract topics from a large unstructured collection of data. Topics can be defined as the main themes that pervade a collection of data or a list of words that occur in statistically meaningful ways.

A popular algorithm for topic modeling is Latent Dirichlet allocation (LDA) (Blei et al., 2003) combined with collapsed Gibbs sampling (Steyvers and Griffiths, 2007). LDA is a generative probabilistic model for collections of discrete data such as texts. LDA uses unsupervised learning since topic models do not require any prior labeling of the documents. It can learn or infer the distribution over topics for each document in a collection as well as the probability distribution over words associated with each topic.

Collapsed Gibbs sampling works as follows. It first iterates over each doc d in the collection of documents D and randomly assigns each word in the document d to one of the K topics. Such an assignment provides both (i) topic representations of all the documents and word distributions of all the topics (although not very good ones). To im-

prove these representations, the model iterates through each word w from each document d to gather statistics about two things: the distribution of topics over words and the distribution of documents over topics. Thus, for each topic t, the algorithm computes two values: (i) the proportion of words in d currently assigned to t, i.e. $p(topic\ t|document\ d)$ and (ii) the proportion of assignments to t over all docs coming from w, i.e. $p(word\ w|topic\ t)$. In a third step, w is reassigned a new topic t selected based on the probability: $p(topic\ t|document\ d) * p(word\ w|topic\ t)$. Finally, the model repeats the previous step a large number of times, until it eventually reaches a roughly steady state where its assignments are quite good. These assignments can then be used to estimate the topic mixtures of each document and the words associated to each topic. Using LDA, we could successfully extract the most salient topics in the Wolof monolingual corpus, as shown in Table 2. As one can observed, the words that make up topics #1 and #4 seem to be associated with the religious domain. Topic #3 is related to the *legal* domain (e.g. the Senegalese constitution) and topic #5 is related to a more general domain, including women's rights.

Topic #	Examples of words related to the topic
1.	*Yàlla* "God", *Yeesu* "Jesus", *boroom* "master", *nit* "people", *kirist* "Christ", *àddina* "world", *yawut* "jews" *gëm* "believe", *bàkkaar* "sins"
2.	*xeet* "race", *olokost* "holocaust", *yawut* "jews", *faagaagal* "murder", *raafal-xeet* "extermination", *cosaan* "tradition", *nguur* "government"
3.	*yoon* "law", *dépite* "deputy", *tànn* "elect", *pénc* "assembly", *sañ-sañ* "authority", *sàrt* "charter", *askan* "nation", *nguur* "government"
4.	*Aji-sax* "the Lord", *israyil* "Israel", *buur* "king", *Daawuda* "David", *Musaa* "Moses", *Misra* "Egypt", *sarxalkat* "priest", *saraxu* "to beg"
5.	*yelleef* "rights", *jigéen* "women", *jàmm* "peace", *farañse* "french", *bokk-réew* "democracy"

Table 2: Most salient topics in the Wolof corpus.

5. Applications

As a first step towards building a French-Wolof NMT system, we used the corpus to create word embedding models (Lo et al., 2019), as outlined in section 5.1.. Section 5.2. briefly presents the LSTM-based models we are currently developing to translate between French and Wolof.

5.1. Wolof Word Embeddings

Basically, a word embedding is a representation of a word as a vector of numeric values. To develop neural word embedding models for Wolof, we used the monolingual Wolof corpus as training data (Lo et al., 2020). At that time, the corpus contained 47457 phrases and a total of 867951 repeated words. In addition, there were 24232 unique words in the vocabulary (only 33% of these occurred more than five times).

To assess the quality of the corpus, we trained three word embedding models on the Wolof monolingual corpus: a continuous bag-of-words (CBOW) model, a Skip-gram model (Mikolov et al., 2013) and a Global vector for word representation (GloVe) (Pennington et al., 2014). The evaluation shows that the models capture word semantic relatedness despite the moderate corpus size. Tables 3, 4 and 5 display, for each word (in the column $word$) its n nearest neighbours, as generated by the CBOW, Skip-gram and the GloVe models, respectively. Please note that, in these tables, the English translation (non bold) is just indicative. Our models only exploit Wolof words occurrences.

word	n_1	n_2	n_3	n_4
afrig africa	**patiriis** patrice	**kongo** Congo	**lumumbaa** lumumba	**reyee** killed
bànk bank	**leble** to lend	**leb** borrow	**cfa** cfa	**koppar** money
banaana banana	**xollitu** peel	**rattax** slippy	**roose** to water	**kemb** peanut
aajo need	**fajug** resolve	**regg** sate	**mbaax** kindness	**solaay** clothing
bamba bamba	**barke** grace	**maam** grand-pa	**ibra** ibra	**seex** sheikh

Table 3: Examples of Wolof words (in bold) with their five nearest neighbours according to CBOW.

word	n_1	n_2	n_3	n_4
afrig africa	**oseyaani** oceania	**asi** asia	**saalumu** south	**sowwu** west
bànk bank	**dugal** put in	**kont** account	**jàngi** go to school	**monjaal** world -wide
banaana banana	**soraas** orange	**màngo** mango	**guava** guava	**xollitu** peel of
aajo need	**fajug** resolution	**aajowoo** want	**faj** to resolve	**faji** resolve
bàmba bamba	**matub** complete -ness	**taalubey** student	**lumumbaa** lumumba	**seex** Sheikh

Table 4: Examples of Wolof words (in bold) with their five nearest neighbours according to skip-gram.

We qualitatively verified the validity of our models by training GloVe on a large-scale French corpus consisting of 350000 Wikipedia French articles. The results of this experiment also indicated similar patterns, as shown in Table 6. Words and their nearest neighbours are semantically related. For example, the three first neighbours of *uranus* (uranus in English), are *jupiter* (jupyter), *saturne* (saturn), and *pluton* (pluto). In Table 6, the first column shows the target words and the three other columns give the first, second, and third nearest neighbours, respectively.

Table 7 gives the results of a qualitative comparison (a similarity method like the word analogy task (Lo et al., 2019)) of the three word embeddings models for Wolof.

An evaluation of the three models indicated that GloVe and Skip-gram give acceptable performance despite the lack of data. The best results are obtained when using Glove.

word	n_1	n_2	n_3	n_4
afrig africa	**oseyaani** oceania	**asi** asia	**gànnaaru** north	**sowwu** south
bànk bank	**fmi** imf	**kont** account	**nafa** wallet	**monjaal** world -wide
banaana banana	**soraas** orange	**màngo** mango	**guyaab** guava	**xob** leaf
aajo need	**fajug** resolve	**tekki** mean	**faju** resolved	**lew** legal
bàmba bamba	**xaadimu** Khadim	**rasuul** prophet	**coloniales** colonial	**seex** sheikh
bant wood	**daaj** press in	**daajoonin** pressed in	**daajee** press with	**daaje** press with

Table 5: Examples of Wolof words (in bold) with their five nearest neighbours according to GloVe.

target word	n_1	n_2	n_3
atom	atomes	isotope	cathode
mathématique	mathematiques	axiomatique	probabilites
art	contemporain	deco	abstrait
peinture	figurative	picturaux	picturales
agriculture	arboriculture	cerealieres	cerealiere
boudhisme	hindouisme	brahmanisme	jainisme
uranus	jupiter	saturne	pluton
planete	extraterrestre	lointaine	orbitant
mer	caspienne	baltique	ocean
fleuve	baikal	fleuves	embouchure

Table 6: French Wikipedia GloVe words neighbours.

couple	CBOW	SG	GloVe
(senegaal, dakaar)	1	0	1
(faraas, pari)	0	0	0
(janq, waxambaane)	0	0	1
(jigéen, góor)	0	0	0
(yaay, baay)	1	1	0
(jëkkër, jabar)	0	0	0
(rafet, taaru)	1	1	1
(teey, yem)	1	1	1
(tàmbale, sumb)	1	1	1
(metit, naqar)	0	0	1
(suux, diig)	1	1	1
(xam, xami)	0	1	1
(ajoor, kajoor)	0	1	1
(taarix, cosaan)	1	1	1
(jàng, jàngale)	0	0	1
Total	47%	53%	73%

Table 7: Scores.

5.2. French-Wolof Machine Translation

In addition to developing word embedding models, we used the corpus to train and evaluate four LSTM based models to translate French sentences into their Wolof counterparts: a baseline LTSM, a bidirectional LTSM, a baseline LTSM + attention, a bidirectional LSTM + attention (Hochreiter and Schmidhuber, 1997; Sutskever et al., 2014b). The models are still under development.

LSTM networks are used for both the encoding and decoding phases. As a first step, the encoder reads the entire input sequence from the source language and encodes it to a fixed-length internal representation. The word embeddings are built using an embedding layer whose dimension is equal to the size of the source language vocabulary. In a second step, a decoder network uses this internal representation to predict the target sentence. Starting from the start of sequence <SOS> symbol, it outputs words until the end of sequence <EOS> token is reached. In other words, the decoder makes prediction by combining information from the thought vector and the previous time step to generate the target sentence.

The model is trained on a dataset of about **70,000** sentences split into training (50%) and validation (50%). The training parameters currently used for the baseline model are displayed in Table 8.

Parameters	Values
Embedding dimension	128
Number of units	300
Learning rate	0.001
Dropout rate	0.25
Number of epochs	500
Batch size	64

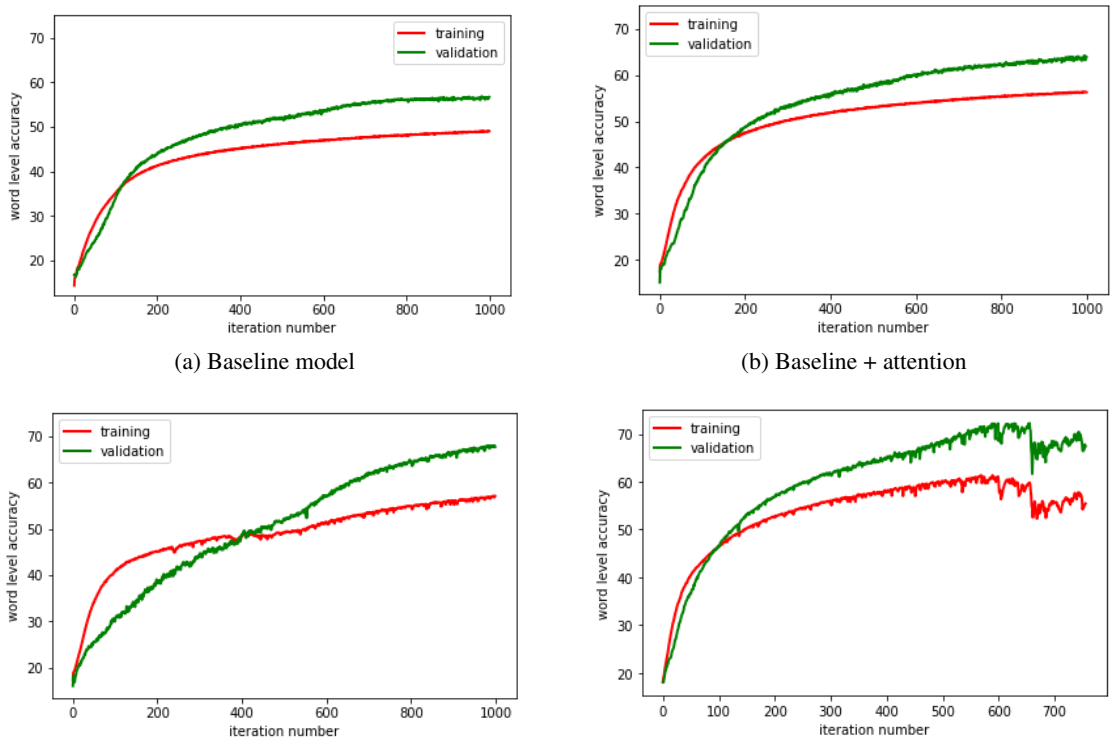
Table 8: Training parameters for the LSTM models.

Across experiments, the following hyper-parameters are kept constant: number of LSTM units, embedding size, weight decay, dropout rate, shuffle size, batch size, learning rate, max gradient norm, optimizer, number of epochs and early stopping patience. All models are composed of a single LSTM layer with a dropout layer for the decoder, dropout rate and weight decay regularization parameters for both the encoder and decoder. Models are trained using Adam stochastic gradient descent with a learning rate set to 10^{-3} (Kingma and Ba, 2014).

Figure 4 shows the current results obtained by the four models in terms of accuracy on the validation set.

As we can see, with local attention, we achieve a significant gain of 7% validation accuracy over the baseline unidirectional non-attentional system. In turn, the unidirectional attentional model slightly underperforms the bidirectional non-attentional model. The best accuracy score is achieved when combining bidirectional LSTMs with the attention mechanism. An accuracy gain of 15,58% could be observed when comparing the latter model with the unidirectional non-attentional baseline.

The current experiments show that there are many opportunities to tune our models and lift the skill of the translations. For instance, we plan to expand the encoder and the decoder models with additional layers and train for more epochs. This can provide more representational capacity for the model. We are also trying to extend the dataset used to fit our models to 200,000 phrases or more. Furthermore, refining the vocabulary by using subword representations such as BPE (byte pair encoding), which have become a popular choice to achieve open-vocabulary trans-



(a) Baseline model

(b) Baseline + attention

(c) Bidirectional without attention

(d) Bidirection with attention

Figure 4: Validation accuracy of the LSTM models.

Our NMT models	Accuracy
Baseline	56.69
+ attention	63.89
+ bidirectional	68.03
+ bidirectional + attention	72.27

Table 9: The performance of the NMT system on French to Wolof dataset. Scores are given in terms of accuracy on the validation. All values are in percentage.

lation. Previous work (Sennrich et al., 2016) has demonstrated that low-resource NMT is very sensitive to hyper-parameters such as BPE vocabulary size. Likewise, recent work (Qi et al., 2018) has shown that pre-trained word embeddings are very effective, particularly in low-resource scenarios, allowing for a better encoding of the source sentences.

6. Conclusion

In this paper, we reported on a relatively large French-Wolof parallel corpus. To the best of our knowledge, this is the largest parallel text data ever reported for the Wolof language. French was chosen particularly because, as the official language of Senegal (the country of the most Wolof speakers), it is easier to find parallel data between French and Wolof than between e.g. English and Wolof.

The corpus is primarily designed for neural machine translation research, but in a way to also satisfy the need for human users. The corpus currently consists of six major domains and is still under development. We are trying to extend it further with material that can be made freely available. Indeed, our plan is to make the parallel corpus publicly available. We are still harvesting more data and we also need to first clarify copyright and licensing issues. In our first experimentation with the corpus, we obtained relatively good results, indicating that the corpus is quite suitable for the development of word embeddings. This also provides a good starting point for further research. Future studies will explore in more details the suitability of the corpus for the development of neural machine translation systems to map Western languages to local Senegalese languages like Wolof and Fula. This paper has only focused on French and Wolof, as our corpus currently mainly contains resources for these two languages. However, we believe that the model described here will still be applicable to the other languages (e.g. Fula and Bambara) that we seek to promote.

7. Bibliographical References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Cissé, M. (1998). *Dictionnaire français-wolof*. Langues & mondes/L'Asiathèque.

Dione, C. M. B. (2014). LFG parse disambiguation for Wolof. *Journal of Language Modelling*, 2(1):105–165.

Dione, C. B. (2019). Developing Universal Dependencies for Wolof. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 12–23, Paris, France. Association for Computational Linguistics.

Diop, B. B., Wülfing-Leckie, V., and Diop, E. H. M. (2016). *Doomi Golo – The Hidden Notebooks*. Michigan State University Press.

Diop, B. B. (2003). *Doomi Golo: Nettali*. Editions Papyrus Afrique.

Fung, P. and Church, K. W. (1994). K-vec: A new approach for aligning parallel texts. In *Proceedings of the 15th conference on Computational linguistics-Volume 2*, pages 1096–1102. Association for Computational Linguistics.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*.

internationale de la Francophonie, O. (2014). La langue française dans le monde. *Paris, Nathan*.

Kesteloot, L. and Dieng, B. (1989). *Du tieddo au talibé*, volume 2. Editions Présence Africaine.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *the tenth Machine Translation Summit*, volume 5, pages 79–86, Thailand, AAMT, AAMT.

Lamraoui, F. and Langlais, P. (2013). Yet another fast, robust and open source sentence aligner. time to reconsider sentence alignment. *XIV Machine Translation Summit*.

Langlais, P., Simard, M., and Véronis, J. (1998). Methods and practical issues in evaluating alignment techniques. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 711–717. Association for Computational Linguistics.

Lilyan, K. and Cherif, M. (1983). Contes et mythes wolof.

Lo, A., Ba, S., Nguer, E. H. M., and Lo, M. (2019). Neural words embedding: Wolof language case. In *IREHI19*.

Lo, A., Dione, C. M. B., Nguer, E. M., Ba, S. O., and Lo, M. (2020). Building word representations for wolof using neural networks. *Springer*.

Ma, X. (2006). Champollion: A robust parallel text sentence aligner. In *LREC*, pages 489–492.

Mangeot, M. and Enguehard, C. (2013). Des dictionnaires éditoriaux aux représentations xml standardisées.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1532–1543.

Qi, Y., Sachan, D., Felix, M., Padmanabhan, S., and Neubig, G. (2018). When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 NAACL-HLT Conference, Volume 2 (Short Papers)*, pages 529–535. Association for Computational Linguistics.

Resnik, P., Olsen, M. B., and Diab, M. (1999). The bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1-2):129–153.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of ACL (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014a). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014b). Sequence to sequence learning with neural networks. *Advances in neural information processing systems (NIPS)*, 18:1527–1554.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *In Proceedings of the RANLP 2005*, pages 590–596.