CHAPTER ONE

INTRODUCTION

1.1 General Preamble

A linguist will wonder if it is possible for any machine to translate a stretch of language without any linguistic knowledge. It is the claim of Statistical Machine Translation (SMT) experts that a machine can translate natural human language without reference to linguistic knowledge. In other words, to develop SMT (one of the two main approaches of Machine Translation), there is no need for linguistic rules. Although, Hassan (2009) observes that Phrase-based SMT[1] lacks the capacity to produce grammatical translations and is deficient in handling long-range reordering while maintaining the grammatical structure of translation outputs, he integrates syntactic structures into the system to produce more fluent MT output. In addition, he is of the opinion that syntax can help Phrase-based SMT systems to produce well-formed translation outputs by the use of syntactically-guided translation, language models and reordering techniques.

It would however be paradoxical if Machine Translation (MT) can be built without recourse to linguistic rule or knowledge. While some are the proponents of the idea that MT is achievable without reference to any linguistic guidance or rules, some others see it as a process of merely modelling human natural language which should include linguistic rules so as to overcome some of the deficiencies of machine in the manipulation of a human language.

Awobuluyi (2010) views MT as a major contribution of linguistics to Information and Communication Technology. He stresses the fact that one of the significant contributions of linguistics to technology, especially Information and Communication Technology (ICT) is the MT. He states that:

> …another operation which researchers would like computer to be able to perform… That operation is known as machine translation, and as its name implies, it involves getting computers to translate well-formed and fully idiomatic written expressions in one language into well-formed and equally idiomatic corresponding expression in another language… (Awobuluyi 2010: 34-35)

[1] Phrase Based SMT is a type of SMT which will be discussed fully in chapter two.

He equally points out some of the limitations the computer encounters when performing the mental activities of humans. He notes that the computer's judgement on human language may be patently incorrect. To him, this shows how difficult it is for now to get a machine to accurately replicate all mental operations that human beings perform. Awobuluyi's view is in line with Hutchins and Somers' (1992) earlier claim.

Hutchins and Somers (1992) claim that the mechanisation of translation has been one of humanity's oldest dreams which came into reality in the twentieth century in the form of computer programmes capable of translating a wide variety of texts from one natural language into another. Hutchins and Somers (1992) maintain that there are no 'translating machines' which, at the touch of a few buttons, can take any text in any language and produce a perfect translation in any other language without human intervention or assistance. They conclude that the ideal that a machine will translate natural human language like humans is for the distant future, if it is even achievable in principle, which many doubt.

As limited as the judgement of the computer could be on human language, the development of MT in African languages, and in particular, Yorùbá language has just begun. Very few Yorùbá-English rule-based and, recently, a few hybrid MTs are available. No SMT is available in the language (except the Google Translate). SMT is said to be the most dominant paradigm in machine translation today because of its accuracy, computational efficiency and fast adaptability to new languages and domains (Lopez 2010).

It is against this background that this research is conducted to find out to what extent a computer can acquire and translate African languages, especially the Yorùbá language and the challenges associated with this efforts since there are enormous works on the English language with other languages like French, German, and Chinese.

SMT relies on large volumes of parallel human translated materials of the language pairs before seemingly acceptable translation could be achieved. There are a very few translated materials for African languages. This may be because African languages are said to be resource-scarce languages from technological point of view (see De Pauw, Wagacha and Schryver, 2011). This indicates that African languages are languages with small or economically disadvantaged users and are typically ignored by the commercial world (Chan and Rosenfeld 2012). But the era of learning English before accessing ICT applications is over. Localisation of ICT is gaining ground with the use of MT as part of its reliable efforts.

Though MT works better for specialised and narrow domains like hotel reservation, flight booking, safety instructions and weather forecast than the general domain because of their restricted register (see Hutchins and Somers1992 and Koehn 2009). It is observed that translations for these specialised domains are scarce at least for Nigerian languages. Egbokhare (2011) observes this and particularly the need for local airplanes to use Nigerian languages in Nigeria. He comments:

> One has heard it said that there are too many languages in Nigeria; hence, it will be virtually impractical to meet the needs of every group. This argument is specious because with five languages, English, Nigerian Pidgin, Hausa, Yorùbá and Igbo, the linguistic needs of over 90% of Nigerians can be met. Nigeria must insist that airlines address Nigerians in the languages they understand best and planes flying in our airspace must adhere to language requirement as part of airline safety requirement…

However, there are literary materials that have been translated basically for academic and social purposes. Then, the question is, to what extent will these literary translations assist to build a SMT for unidirectional Yorùbá-English language translators?

Yorùbá language is a tone language like most other African languages. Tone performs phonological, syntactic and semantic functions which are necessary for accurate and perfect translation. However, *Moses* which is one of the most used open source platforms to build SMT systems is widely used on Mac and Linux. The concern, therefore, is that to what extent will Moses accommodate the tone nature of Yorùbá language in the translation of Yorùbá texts?

The thrust of this thesis is not just to build SMT for Yorùbá-English language pair but to subject its output to syntactic analysis so as to make contribution to the development of SMT and Moses in particular from the perspective of African languages.

1.2 Machine Translation (MT)

Machine translation (MT) is an automated translation. It is the process by which computer software is used to translate a text from one natural language (such as English) to another (such as Yorùbá). Odoje (2010:5), quoting (Reifler 1954) perceives, it as a complete mechanised process without pre- or post-editor’s intervention, the output of which must be satisfactory with regard to both semantic accuracy and intelligibility. Taking into account semantic accuracy and intelligibility, translation is not an easy task both for human and machine because no two languages can be mirrored exactly the same way. This informs why Jurafsky and Martin (2000:1) assert that a perfect translation is an illusion because of cultural differences that exist among native speakers of different languages. Hence, some of the semantic contents of the source language vanish during translation. Consider example (1) below in which Àkàrà-oògùn is narrating his father's experience in Fágúnwá's *Igbo-Olódùmarè*. He reports the impact of other creators supporting Olówó-ayé when Olówó-ayé is fighting Igbó-Olódùmarè's gate keeper, Àǹjànnú-ìbẹ̀rù that:

1. …ni wọ́n ti tẹ gọ̀ǹgọ́ mọ́ ojú ìlù …

(Fagunwa 2005b:38)

This text from Fagunwa's *Igbó Olódùmarè* has been translated by Ajadi (2005) and Soyinka (2010). In the translations, it was observed that the poetic beauty in the source language is quite lacking in the translated equivalence where Ajadi and Soyinka tries to ensure they employ appropriate language to represent the poetic beauty in the target language as seen in (2a and 2b) below:

2a. Ajadi: they began to intensify their praise drumming
2b. Soyinka: drumsticks dug into drumskin, intoning

In spite of some missing poetic beauty (aesthetics), the translators are able to achieve this remarkable translation because both of them have the mastery of the languages (English and Yorùbá) which the computer does not, in a sense possess. In fact, the computer does not understand any natural language, rather it models it. Hence, no one should expect the computer to translate like humans. As earlier mentioned, no 'translating machines' at the touch of a few buttons, can take any text in any language and produce a perfect translation in any other language without human intervention or assistance.

Another point worthy of note is that two translators can never translate the same sentence exactly the same way. This is necessitated by the translator's language experience and choice. Hutchins (2001:5) reiterated Holmstrom's definition of translation, which takes into consideration the educational qualification and personality of the translator.

> Translation is an art; something which at every step involves personal choice between uncodifiable alternatives; not merely direct substitutions of equated sets of symbols but choices of values dependent for their soundness on the whole antecedent education and personality of the translator.

This is observed in Ajadi (2005) and Soyinka (2010) translations of Fagunwa's *Igbo-Olódùmarè*. Consider example (3) below:

3. Gba eléyìí, jẹun dáadáa, má ṣe jẹ́ kí inú run ọ́, ọkọ kìí ju ọkọ lọ.

4a. Ajadi: "Take this, eat very well, and try to avoid stomach ache; one husband does not surpass the other

4b. Soyinka: "Take this, eat soundly, don't let anything upset your stomach, no husband is more treasured than another."

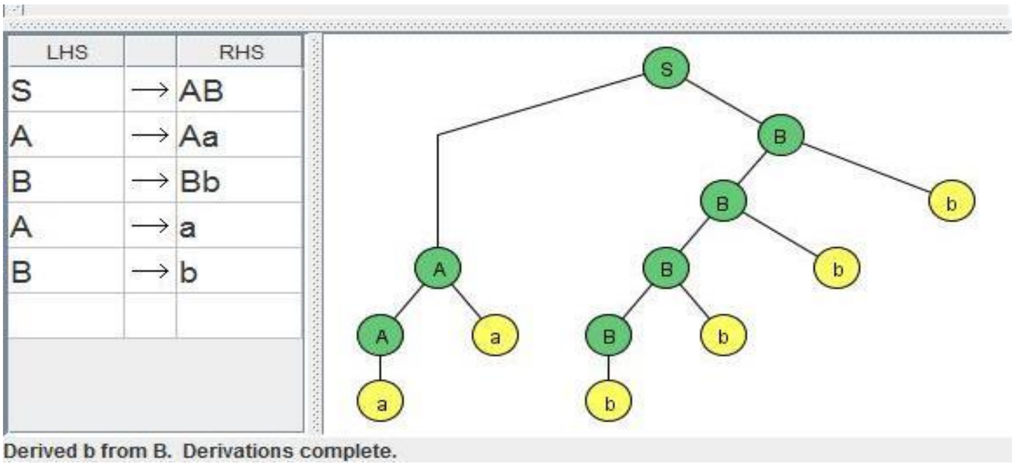
(4a&b) are the translations of (3) though different words, which reflect the personality and choie of words of the translators. The computer has no personality neither does it have choice of words like human translators yet it translates. This raises the question "Can computer translate?"

1.3 Can Machine Translate?

Translation involves mastery or high level of competence in the language pairs. This will assist the translator to take note of minute but very important differences of the languages to be translated. As observed by Osundare (1995), despite pan-human cultural and linguistic traits, each culture as well as the language in which it is articulated, has a certain degree of uniqueness. This language uniqueness is to be captured in translation processes before real translation could be achieved.

If translation involves this high level of human language capacity, how then can a machine which does not have human language competence translate a human language with its linguistic and cultural uniqueness? However, the answer to the question at hand is yes! After all, MT is easily accessible online although the translation may not be perfect. For example, Google Translate translates *I love Nigeria* to Spanish as: "*Amo Nigeria";* German as: *Ich liebe Nigeria* and to Yorùbá as *Mo ní ìfẹ́ Nigeria* as well as Arabic: أنا أحب نيجيريا. But on another consideration, it could be concluded that no machine or computer has human language intuitiveness for translation. In fact, computer has no language except programming language. Hence the machine becomes a tool adopted for translation.

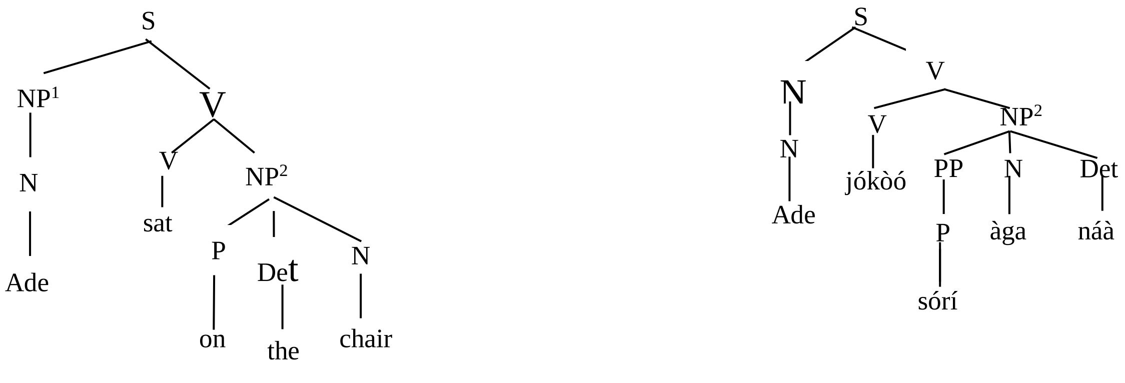
To a computer scientist, language is a set of string (Sipser 2006:14). It has its grammar and the string can be parsed into syntactic tree as it is done with natural language. Consider the tree below.



LHS		RHS
S	→	AB
A	→	Aa
B	→	Bb
A	→	a
B	→	b

Derived b from B. Derivations complete.

In the diagram above, the language is aabbbb and it is parsed as shown by the structure above. It is this kind of structure that forms the base for Rule-Based MT where the simplified structure of a language like a Phrase Structure Rule is compiled into computer language programming for translation. Consider Eludiora, Salawu, Odejobi and Agbeyangi's (2011) syntactic tree diagram used for Rule-Base MT below:



You will observe that just as programming language structure is labeled so is the natural language. However, the syntactic structures of the natural language above are misinforming in that an N with or without its compliment should project to an NP. In our opinion, the NP^2 should be a PP headed by a P. But as observed in the structure above, P and PP respectively project to NP^2 which shows inconsistence in the analysis and the claims of the structure as well as the realities of the languages syntactically.

However, the concern of this work is that Rule-Based MT like the above uses the syntactic structure of languages involved for the purpose of translation. Therefore, the translator is not the computer but the person who programmed it for translation. The mastery of the languages involved as well as the computer programming language adopted by the programmer will determine the performance of the machine among other factors.

Concluding that machine does not have an input in the MT process is basically limiting the scope of MT to Rule Based approach and machine rule learning. Primarily, there are two ways a machine learns: rule learning and statistical learning. Rule learning requires specific rule or pattern for computer to learn and generalize like Rule Based MT. This implies that a specific rule of operation is given to the computer in order to perform a task. The challenge with this is that, traditionally, programming a computer is to tell a computer what to do. This is represented in numbers and expressed in formulas. Sometimes, problems or tasks are difficult to express in formula. The way out is for the machine to learn the process of carrying out such tasks by itself; hence, the need for statistical machine learning. Statistical machine learning is a process where a machine learns the process of classifying or carrying out a task by itself from the corpus; the more the corpus the better the result of the learning. There are two ways to do this: supervised learning and unsupervised learning.

Supervised machine learning comprises of algorithms that are generated from externally supplied instances to produce general hypotheses which then serve as predictions for future instances. Generally, with supervised learning there is a presence of the outcome variable to guide the learning process (see Omary and Mtenzi 2010). In other words, supervised learning involves the intervention of humans in the process of learning which may be very expensive most especially when the specifications are much with the consideration of a very large corpus. On the other hand, the process where no human intervention is required is called unsupervised learning. The computer figures out the process and carries out the task by itself. Omary and Mtenzi (2010) explain that unsupervised learning builds models from data without predefined classes or examples. This means no “supervisor” is available and learning must rely on guidance obtained heuristically by the system examining different data or the environment. The output states are defined implicitly by the specific learning algorithm used and built in constraints.

Relating the aforementioned to machine translation therefore both human and the computer are involved in the Statistical Machine Translation. While a human does the translation which serves as the data for machine learning process, the computer uses the statistical method of analysis to acquire necessary information to translate natural human language by itself. Conclusively, a machine can translate but humans have more effective translation capacity for machine to learn from.

That is the bedrock for SMT, the quality of SMT depends largely on the quality and quantity of parallel corpus of language pairs (Hutchins and Somers 1992**,** Frederking and Taylor 2004, Wilks 2009 and Koehn 2010). Therefore, MT is not in competition with human translators (as some are thinking) rather a compliment (aid) for both professional and non-professional translators.

1.4 Brief History of Machine Translation

Writers on the history of Machine Translation have two broad approaches. Some trace the history to the pre-computer era while others begin their historical trace to the advent of the electronic computer. Hutchins and Somers (1992:5) report that the use of mechanical dictionaries to overcome barriers of language was first suggested in the 17th century. They explain that both Descartes and Leibniz speculated the creation of dictionaries based on universal numerical codes. Some examples were published in the middle of the century by Cave Beck, Athanasius Kircher and Johann Becher. The inspiration was the ‘universal language’ movement, the idea of creating an unambiguous language based on logical principles and iconic symbols (as the Chinese characters were believed to be), with which all humanity could communicate without any fear of misunderstanding. The most familiar approach is the interlingual which was elaborated by John Wilkins in his *Essay towards a Real Character and a Philosophical Language* (1668).

Subsequently there were many more proposals for international languages (with Esperanto as the best known) but few attempts were made to mechanise translation until the middle of the century. Prominent among those who made attempt to develop MT are Georges Artsrouni (a French-Armenian) and Petr Troyanskii (a Russian) who applied for patents for ‘translating machines’ in the mid-1930s. Of the two, Troyanskii's was the more significant, proposing not only a method for an automatic bilingual dictionary but also a scheme for coding interlingual grammatical roles and an outline of how analysis and synthesis might work. However, Troyanskii’s ideas were not known until the end of the 1950s. Before then, the computer had been born (see Hutchins and Somers 1992, Hutchins 2001, Hutchins 2007).

Koehn (2010:15) is of the opinion that efforts to build MT systems started almost as soon as the electronic computers came into existence. The prominent aim of developing MT then was to decode messages from the then world powers (Britain and USSR). For example, Koehn (2010:15) reports that Britain used the computer to crack the German Enigma Code in World War II and decoding language codes seemed like an apt metaphor for machine translation. He quotes Warren Weaver who is the first MT researcher to have said:

> When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode (Weaver, 1947, 1949 in Koehn 2010:15).

Hutchins and Somers (1992) are of the view that both Warren Weaver of the Rockefeller Foundation and Andrew D. Booth, a British crystallographer were the first set of people to discuss the possibility of using the computer for translation and they both collaborated with Richard H. Richens (Cambridge), who had independently been using punched cards to produce crude word-for-word translations of scientific abstracts. However, it was Weaver's memorandum of July 1947 that brought MT to the general notice. Research began in various centers in the United State in 1951 and a full time researcher Yehoshua Bar-Hillel at MIT was appointed the same year. In 1952, Yehoshua Bar-Hillel convened the first MT conference where the outlines of future research were discussed. There were proposals for dealing with syntax, suggestions that texts should be written in controlled languages, arguments for the construction of sublanguage systems, and recognition of the need for human assistance (pre and post-editing) until fully automatic translation could be achieved.

In January 1954, there was the first public demonstration of an MT system which is the handiwork of Leon Dostert at Georgetown University who collaborated with IBM. Some Russian sample sentences were carefully selected which were translated into English, using a very restricted vocabulary of 250 words and just six grammar rules. Although it had little scientific value, it was sufficiently impressive to stimulate the large-scale funding of MT research in the United States and to inspire the initiation of MT projects elsewhere in the world, notably, in the Soviet Union. (See Slocum 1985, Hutchins 1995, Hutchins 2007, Koehn 2010, Chéragui 2012).

Optimism was very high to develop a full automated machine that will translate human language and probably replace human translators. As a result of this, many universities across the globe were involved in the development of MT. But disillusionment grew as the complexity of the linguistic problems became more and more apparent. In a 1960 review of MT progress, Bar-Hillel criticises the prevailing assumption that the goal of MT research should be the creation of Fully Automatic High Quality Translation (FAHQT) systems producing results indistinguishable from those of human translators. He argues that the 'semantic barriers' to MT could in principle only be overcome by the inclusion of vast amounts of encyclopaedic knowledge about the 'real world'. His recommendation was that MT should adopt less ambitious goals; it should build systems which will make cost-effective use of human-machine interaction.

Automatic Language Processing Advisory Committee (ALPAC) was set up in 1964 and their report was given in 1966. The report states that MT was slower, less accurate and twice expensive compared to human translation and that there was no immediate or predictable prospect of useful Machine Translation. The report put on hold funding in respect to the activities of MT in US for twenty years. However, research continued in other parts of the world. The revival of MT research in the 1980s and the emergence of MT systems in the marketplace have led to growing public awareness of the importance of translation tools. There may still be many misconceptions about what has been achieved and what may be possible in the future but the healthy state of MT is reflected in the multiplicity of system types and of research designs which are now being explored and which were undreamt of when MT was first proposed in the 1940s. Further advances in computer technology, Artificial Intelligence and theoretical linguistics suggest possible future lines of investigation while different MT user profiles (the writer who wants to compose a text in an unknown language is a possibility) lead to new designs. But the most fundamental problems of computer-based translation are concerned not with technology but with language, meaning, understanding, and the social and cultural differences of human communication (see Hutchins and Somers 1992, and Hutchins 1994).

According to Koehn (2010) SMT systems are currently being developed in a large number of academic and commercial research laboratories. Some of these efforts have led to the establishment of new companies. Language Weaver was the first company founded in 2002 that fully embraced the new paradigm and promised *translation by numbers*. Commercial statistical machine translation systems are also being developed by large software companies such as IBM, Microsoft, and Google.

1.5 Types of Machine Translation

There are three types of Machine Translation (Slocum 1985:2, Hutchins 1995:1):

- Fully Automatic (automated) Machine Translation (FAMT)
- Machine-aided Translation (MAT),
- Terminology Data bank.

Developing FAMT was the main aim of MT initially. In this type, the text of a language is fed into the computer and the computer automatically produces accepted translation in the target language without any human assistance either at pre or post-editing processes. This aim has not been achieved though work is ongoing and new methods and strategies are adopted to get close to the aim day by day.

MAT systems have two subgroups: Human-Assisted Machine Translation (HAMT) and Machine-Assisted Human Translation (MAHT). HAMT refers to a system where the computer is responsible for producing the translation per se but may need human monitoring at many stages along the way. For example, it could ask a human being to disambiguate a word with regard to its part of speech or meaning, or to indicate where to attach a phrase, or to choose a translation for a word or phrase from among several options discovered in the system's dictionary. MAHT refers to a system whereby a human being is responsible for producing the translation per se but may interact with the computer in certain prescribed situations such as requesting assistance in searching through the dictionary.

A Terminology data bank offers technical terminologies that are usually not common expressions. The main advantage of a terminology data bank may not be automated but that is up-to-date. A technical term is constantly changing and published dictionaries are essentially obsolete by the time they are available. It is possible for a terminology data bank to contain more entries because it can draw on a large group of active contributions or users.

1.6 Processes of Translation

There are two main translation processes: metaphrase and paraphrase. Metaphrase is “Word-to-Word translation (Tripathi and Sarlchel 2010). By implication, it is literal translation, and in most cases, it may not convey semantic meaning of the original text. For instance consider example (5) below:

5. Mi ò lè pa ara mi
I neg can kill myself
Literal translation: I cannot kill myself
Translation: I'm fed up

(5) Above is a statement of frustration but the literal translation did not convey the meaning. Literal translations like this is found in the Nigerian English.

Paraphrase relates to "dynamic equivalence" this means that the translated text would contain elements of the original text but may not necessarily contain the word-to-word translation (Tripathi and Sarlchel 2010) this is seen in (5) above where *Mi ò lè pa ara mi* is translated as *I'm fed up*.

The translation processes are incorporated into MT as would be observed in the various approaches to MT.

1.7 Approaches to Machine Translation

There are three main approaches to MT: dictionary based MT, rule based MT and corpus based MT.

1.7.1 Dictionary Based Machine Translation

Tripathi and Sarlchel (2010) explain that this method is based on entries of language dictionaries whereby equivalent words are used for translated verses. They report that the first generation of MT in the 1940s to the mid-1960s was entirely based on machine readable or electronic dictionaries. This approach may translate word for word and some phrases but will not translate sentences. There are more to sentence translation than word substitution. Most of the other translation approaches utilize bilingual dictionaries with grammar rules.

1.7.2 Rule Based Machine Translation

This is the incorporation of linguistics rules to computer for the translation process. Aside from millions of dictionaries for the language pair; morphological, syntactic and semantic information about the source and the target language is logically adapted to computer algorithm for translation. The methods used in this approach are discussed below using a diagrammatic pyramid triangle adopted from Tripathi and Sarlchel (2010).

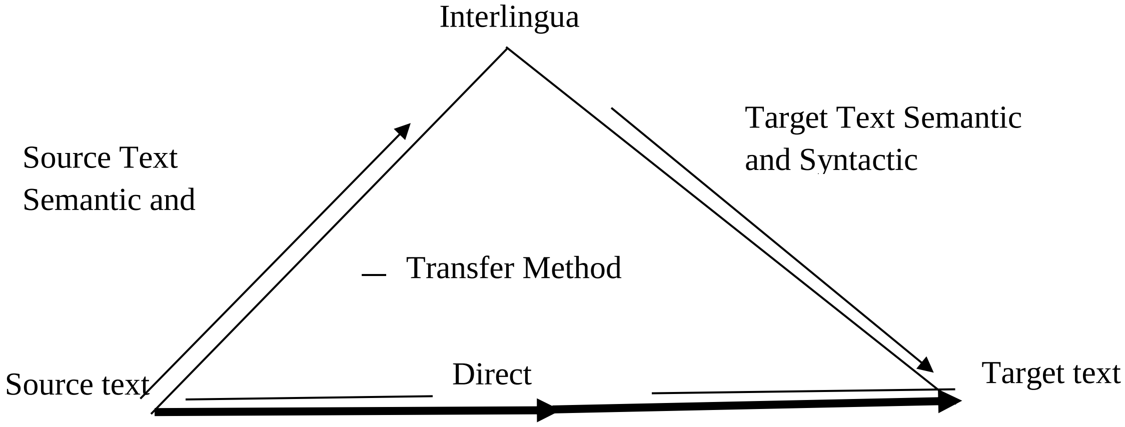


Fig. 1: A Diagram illustrating the Rule Based Approach (Tripathi and Sarlchel 2010)

1.7.2.1 Direct Approach

This is the first generation MT systems. Hutchins and Somers (1992) explain that the strategy lacks any kind of intermediate stage in translation process: the processing of the source language input text leads 'directly' to the desired target language output text. They report that the first generation MT systems were very primitive even in comparison with the lowliest electronic calculators of today because there were no high-level programming languages. Most programmes were done in assembly code. In broad outline, first generation direct MT systems began with what they called a morphological analysis phase where there would be some identification of word endings and reduction of inflected forms to their uninflected basic forms, and the results would be input into a large bilingual dictionary look-up programme. There would be no analysis of syntactic structure or of semantic relationships. In other words, lexical identification would depend on morphological analysis and would lead directly to bilingual dictionary look-ups providing the target languages word equivalences. There would follow some local re-ordering rules to give more acceptable target language output, perhaps moving some adjectives or verb particles, and then the target language text would be produced.

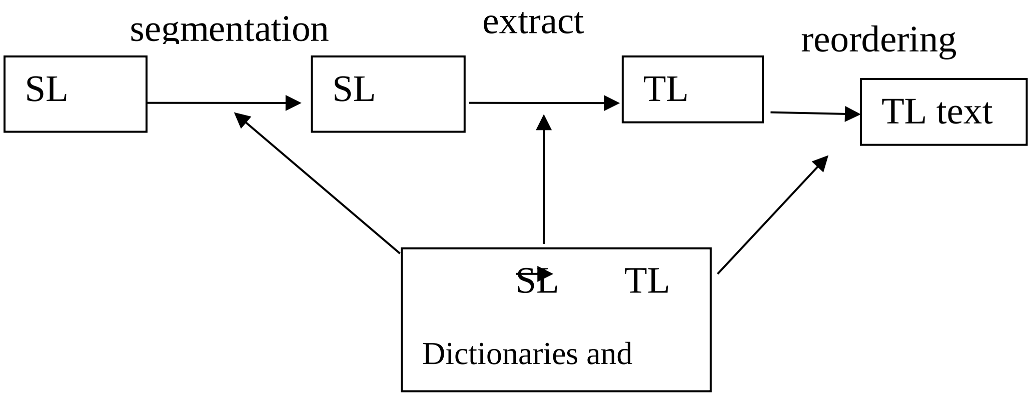


Fig. 2: A Diagram showing the Direct approach Model (adapted from Hutchins 2007)

From the model linguistic structure and relationship between words are not taken into account for translation. This is seen as one of the major setbacks for the strategy. For example, Noone (2003:14) cited mistranslation between Russian and English, using a poem by Rose Saperstein. She then concludes that if this approach is used with other strategies, it may produce better result. The deficiency of direct approach brought about the indirect approach.

1.7.2.2 Transfer approach

This approach operates on the basis of the known structural differences between the source and target language. A transfer system can be broken down into three stages: analysis, transfer, and generation. The transfer method presupposes a parse tree of the input in the source language; this is known as the analysis stage. This parse tree is then mapped to a parse tree of the target language. This means that semantically equivalent but syntactically different trees of the source language are mapped to the target language. After finding the parse tree of the target language, it is put into some grammar module which will take the tree as input and will output the corresponding natural language sentence. When the source language is used to produce a parse tree, this parse tree will usually contain the base form of the words in the sentence instead of their inflected forms. The parse tree will represent the main elements of the sentence, (nouns, verbs, complements) and many sentences of the source language can be represented by the same tree structure but with different lexical words in them. The model is represented as:

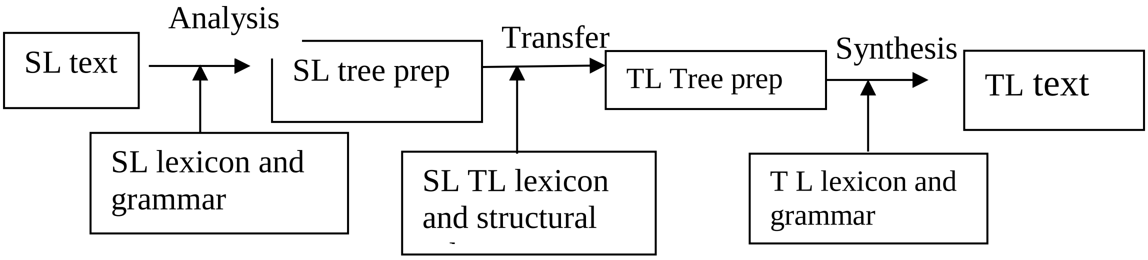


Fig. 3: A diagram showing Transfer Model **(**adapted from Hutchins 2007)

1.7.2.3 Interlingual Approach

The interlingual approach is one of the most attractive for multilingual systems. Each analysis module can be independent, all other analysis modules and all generation modules have no effect on any processes of analysis; the aim of analysis is the derivation of an 'interlingua' representation (Hutchins 2007). Noone (2003) explains that the machine works in a way that a source sentence is analyzed, its semantic content, i.e. meaning, is extracted and represented by an independent language (just a representation or artificial language). This means that a natural language sentence can be generated by using a generation module between the representation of language and the target language. To include an additional language to the translator of this type, we need to simply add analysis module and a generation module for the new language to be represented. An interlingual system can translate between all pairs of language that are represented because it offers the advantage of a system that grows linearly, 2n, where ‘n’ is the number of languages. For example, if the system had 6 modules which are:

Yorùbá – Analysis Generation
Igbo- Analysis Generation
Hausa- Analysis Generation

The machine could translate in all directions; it could even translate Yorùbá to Yorùbá using ‘back translation’. Back translation could give back a syntactically different sentence but leave the meaning intact.

As good as this approach is; it has its deficiencies. One of the deficiencies is finding a language independent representation which retains the precise meaning of a sentence in a particular language and could be used to generate a sentence of a different language. This could be a seriously challenging task. Considerations which must be dealt with are the decision of which representational ontology to use and how to store language-specific details in a general representation (Noone, 2003:20). This means that once a sentence is stored in its interlingual representation, there is no need to look back at the source language because all relevant information is stored in its new interlingual form. Two of the projects using this approach are: ULTRA, UNITRAN.

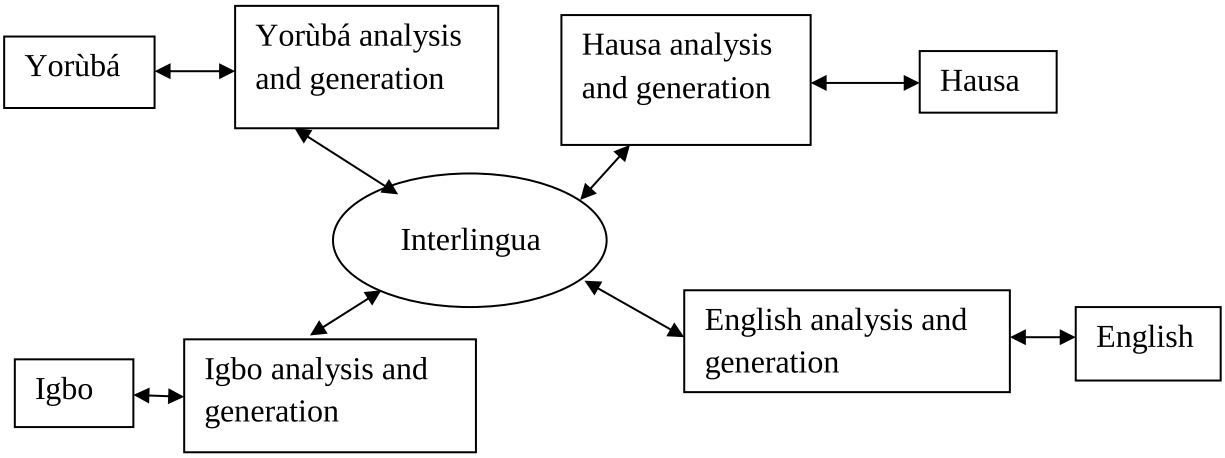

Yorùbá
Yorùbá analysis and generation
Hausa analysis and generation
Hausa
Interlingua
English analysis and generation
English
Igbo
Igbo analysis and generation

Fig. 4: A diagram showing Interlingual Approach Model **(**adapted from Noone 2003 *labelling is mine*).

1.7.3 Corpus Based Approach

This approach is referred to as the empirical approach to MT by Khalilov (2009). Since 1989, corpus approach for machine translation has emerged one of the widely explored areas in machine translation. Because of the high level of accuracy achieved during the translation this method has dominated other approaches. Some of the methods of this approach are: Statistical Machine Translation (SMT), Example Based Machine Translation (EBMT), and Context Based Machine Translation (CBMT) (Tripathi and Sarlchel 2010).

1.7.3.1 Statistical Machine Translation (SMT)

SMT is an approach to MT that is characterized by the use of machine learning methods. It treats the translation of natural languages as a machine learning problem. This means that learning algorithm is applied to a large body of previously translated text known variously as a *parallel corpus*, *parallel text*, *bitext*, or *multitext*. The learner is then able to translate previously unseen sentences (Lopez, 2008).

The system examines many samples of human-produced translation and the SMT algorithms automatically learn how to translate. In less than two decades, SMT has come to dominate academic MT research and has gained a share of the commercial

MT market. Progress is rapid, and the state of the art is a moving target. However, as the field has matured, some common themes have emerged (see Lopez, 2008; Tripathi and Sarlchel 2010; Koehn 2010). There are many language toolkits that could be used for this exercise and the preferable ones are open source tools like Pharaoh, Moses, Joshua, Jane, cdec, phrasal and so on (see http://www.statmt.org/ for details).

It should be noted that this approach can be adapted for both metaphrase and paraphrase translation processes. The system uses a mathematical application of probability for translation. The highest frequency of occurrence is used for translation and the followings are the current approaches to SMT: Phrase-Based Translation, Factored Translation Model, Hierarchical Approach, N-gram-Based Approach, Syntax-Based Approach and so on. Details of these will be discussed in chapter two of this work.

1.7.3.2 Example Based Machine Translation (EBMT)

Guvenir and Cicekli (1998) report that this method is based on analogical reasoning which involves two examples translation. It assumes that a bilingual parallel text exists to derive a translation. This approach offers the advantage of producing results quickly. A further advantage of this approach is that unlike the other methods, this method does not require large-scale knowledge about the source and target languages i.e. grammars, transfer modules, etc. This type of system generally does not deal with analyzing or generating a language's morphology either. Instead, it takes translations without considering what case they are in and stores them just in memory without any change to the representation. This can lead to inaccurate morphological inflection of words in a highly inflectional language.

1.7.4 Hybrid Approach to Machine Translation

Chéragui (2012) explains that some recent works have focused on hybrid approaches that combine the rule-based approach with one of the corpus–based approaches. This was designed to work with fewer amounts of resources and depend on the learning and training of transfer rules. The main idea in this approach is to automatically learn syntactic transfer rules from limited amounts of word-aligned data. This data contains all the needed information for parsing, transfer, and generation of the sentences. The approach seems to favour languages with resource scares since it is a combination of the rule-based and the corpus-based approaches. Its adoption will assist to adjudicate the ambiguity in example (6) below:

6. Ó gbé ìbọn fún ọdẹ
   Sg give gun prep hunter
   He gave the gun to the hunter

(6) Above is merely a literal translation that the rule-based translation can achieve. But (6) also has its figurative equivalence where the semantic content of *gbé* is expanded to mean *to shoot*. This definitely may be difficult to achieve through the rule-based approach (See Odoje 2010) and there is the need for a corpus approach where the system will have come across *gbé* as *to shoot* in its training corpus whereby a weight would have been allotted to it. Also, *ọdẹ* could connote either a hunter or a security guard. Using either of the meaning of *ọdẹ* will still render the translation meaningful. Hence, context is the only option that could demarcate the difference between the meanings of *ọdẹ* as a hunter and a security guard. This necessitates the need for the hybrid approach if the system could capture contextual meanings. Hence, (6) above could be given a translation like (7) below:

7. He shot the security guard

Kuang-Hua Chen and Hsin-His Chen (1996) had earlier stated that this system combines both the advantages and disadvantages of rule based MT and Corpus based MT. They conclude that this integrated approach has good performance and the post-editing efforts needed are very small. But the question remains: to what extent does the system handle hierarchy in its translation? In order words, when does the system resort to translation of figurative expressions like (5 & 6) above? It should be noted that both Kuang-Hua Chen and Hsin-His Chen (1996), and Chéragui (2012) did not state whether the hybrid approach systems can handle contextual meaning which are germane to translation in any respect.

Kuang-Hua Chen and Hsin-His Chen (1996) gave their system translation model as seen below:

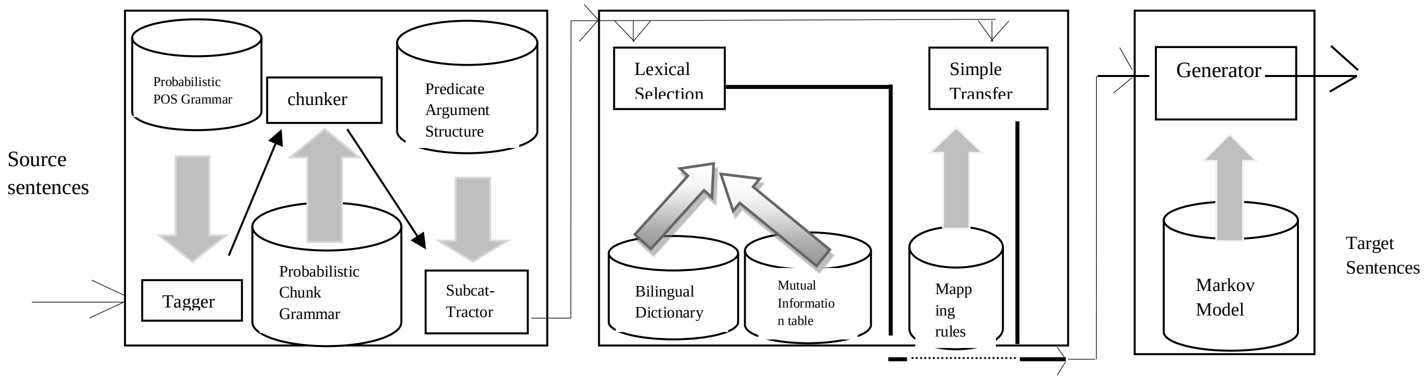


Fig. 5: A diagram showing Hybrid Approach to MT (Adapted from Kuang-Hua Chen and Hsin-His Chen (1996)

1.8 Model Adopted and Reasons for Adopting it

This research adopts phrase based approach to SMT because the researcher is interested in both the formal and natural syntactic analysis of the output of the system so as to contribute to the ongoing discussion on the development of a fully automated machine translation, especially for African languages, since the contribution of the African languages is still very minimal in the general discourse of MT. Moses, which is one of the widely used language toolkit is adopted because it could be easily modified for application to any language.

1.9 Objective of the Study

The aim of the study is to contribute to the ongoing discussion on the Statistical Machine Translation from the perspective of African languages. African languages have some features different from indo-European languages to which Machine Translation has been largely applied. Another significant objective of the study is to build unannotated corpus to train machine statistically for the purpose of translation. It will also be of interest to study how Moses, which is widely used on Mac and Linux accommodate tone languages like Yoruba. This work also aims at studying the extent to which literary texts could help in building a unidirectional statistical machine translation from Yorùbá to English.

1.10 Research Question

This research has five questions to answer. They are as follows:

1. To what extent can the computer acquire and translate African languages especially Yorùbá into English?
2. To what extent will translated literary text assist to build a SMT that is unidirectional Yorùbá-English language translator?
3. To what extent will Moses accommodate the tone nature of Yorùbá language which is very important for translation?

1.11 Significance of the Work

Africa is said to be one of the fastest growing economies in the world. Virtually all African nations along the coast now have crude oil. Almost every electronic and automobile industry is heading towards Africa because of the viability of the market in Africa. This makes the translation of manuals tedious for human translation hence having a MT to serve as an aid will not just assist human translator but will improve the quality of their jobs and reduce the time spent on such jobs.

Aside from the advocacy for the use of indigenous languages for any purpose and in all areas of life championed by Kọ́lá Owólabí which has necessitated the translation of many documents, this research will aid the further campaign for the use of indigenous languages as no language is superior to the other.

Another significance of this study is to contribute to the discussion on MT from the African language perspective so as to make significant contributions to the development of a fully automated machine translation. This study is also important because, it will assist to put on record the process of a machine learning tone languages, a class to which many African languages belong. Lastly, the research would serve as a means of resolving some unresolved arguments about syntactic positions via the technological point of view.

1.12 Limitation of the Study

One of the major limitations of the study is the limited electronic copy of translated materials and the scarcity of a computer system with the required high memory capacity. Hence, the focus is limited to selected literary texts so as to save time and cost. This study will focus on phrase-based approach in Moses Baseline System since it is the starting point to build a SMT system. It is also limited to basic sentence for its test and evaluation hence, proverbs, figurative expressions and so on are not part of it testing and evaluation tools.

1.13 Theoretical Framework

This study adopts Mona Baker’s Corpus-Driven approach also known as Corpus-Based approach to translation studies. Shen (2010:182) explains that Baker was the first scholar to have applied corpus explanation to translation phenomenon in the middle of 1990s. Corpora application to translation research includes mainly parallel, comparable and multilingual corpora (see Baker 1993 in Shen 2010:182). Fernandes (2008:91) reclassifies Baker’s types of corpora to two: parallel and comparable corpora as can be seen in figure (6) below:

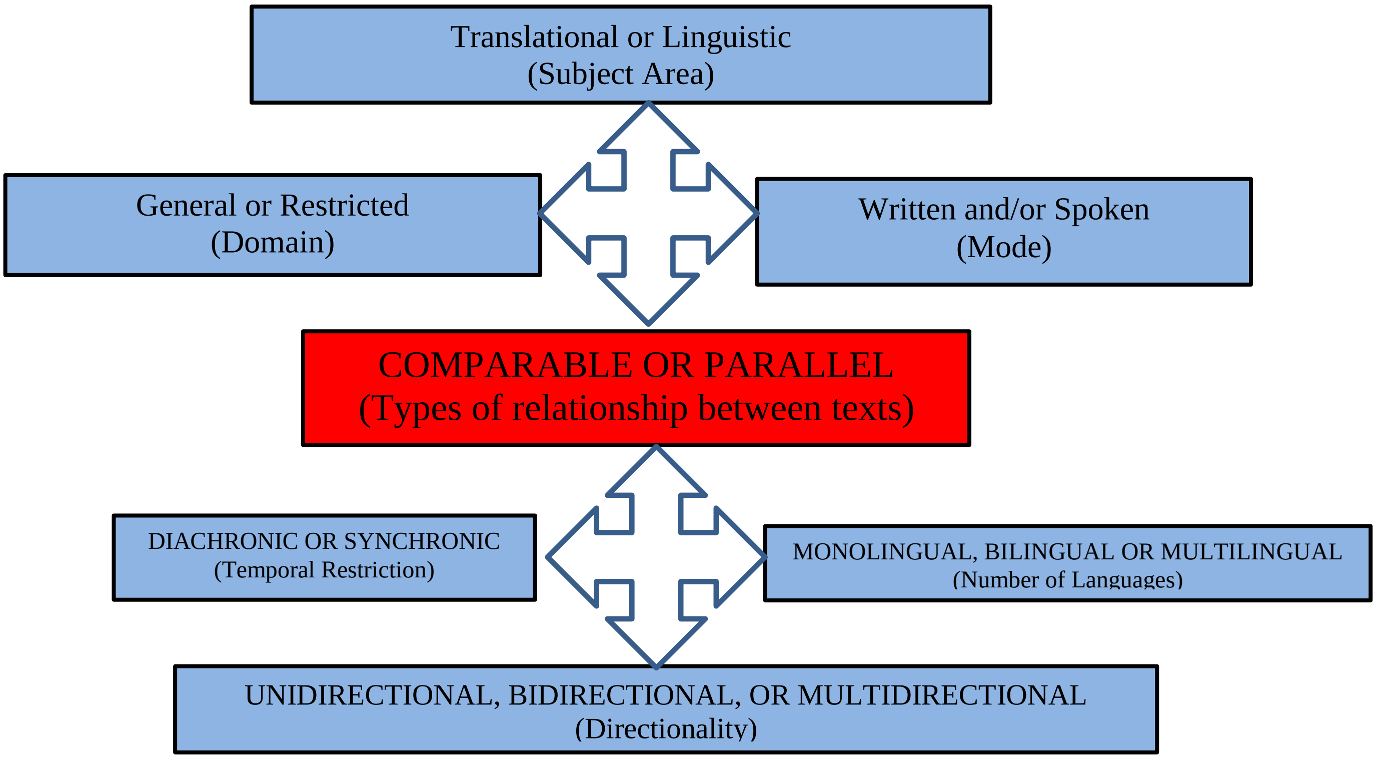


Fig. 6 showing types of corpora adapted from Fernandes (2008:91)

1.13.1 Subject Area: Linguistic or Translational

This Fernandes (2008:92) uses to distinguish between corpus-based studies designed for the study of languages and those built up with the view of investigating translation products and processes. He calls the former linguistic and the later translational. This study is not interested in the translation products and processes but specifically focus on the linguistic aspect; so as to examine the translation of a computer software (Moses).

1.13.2 Domain: General or Specialised

Fernandes (2008:92) refers to the term domain to the area of language enquiry which a corpus focuses. There are two main types of corpora in relation to domain: general or specialised. General in the sense that its content are broader in scope because it is built to study the language of translated material as a whole. By contrast, a specialised corpus looks into the language of specific translated genres or text types. He illustrates Kenny's German-English Parallel Corpus of Literary Text (GEPCOLT) as a specialised corpus which main focus is to investigate the language of translated literary texts from German into English.

1.13.3 Mode: Written and/or Spoken

Mode has to do with the way the original content of a text are delivered. For instance, a text transcribed from an audio or video source is considered “spoken” and a text scanned from a book and converted into electronic form is considered “written” (Fernandes 2008:93). Fernandes (2008:93) notes that there are instances where the text of a corpus can consist of both written and spoken languages. He provides examples like British National Corpus (BNC) which is a general linguistic corpora which comprises of 100 million words collected from samples of written and spoken languages from a wide range of sources.

1.13.4 Temporal Restriction: Diachronic or Synchronic

A corpus can be categorised as synchronic when it focuses on an object of study at one particular point in time. However, when a corpus is concerned with the historical development of an object through time. Fernandes (2008:93) explains that Munday’s (1998) analysis of translation shift is a typical example of a synchronic corpus-based study. Munday small-scale corpus comprised of a short-story by Gabriel Garcia Marques published in Spanish, focuses on the publication year of the English translation. If Munday had decided to include other English translation of the same short story published in different dates – aiming at examining the way these translations changed over time – the study would be of a diachronic kind.

1.13.5 Number of Lnaguages: Monolingual, Bilingual or Multilingual

A corpus can be classified as bilingual if it has two languages and multilingual if it has more than two languages. Fernandes (2008:94) emphasis that another aspect related to the number of languages being represented in the orpus has to do with language varieties. If a corpus is bilingual involving Portuguese and English; it is important to specifiy the language variety of these two languages (i.e European Portuguese vs Brazilian Portuguese and Bristish English vs American English)

1.13.6 Directionality: Unidirectional, Bidirectional or Multidirectional

Zanettin (2000) explains in Fernandes (2008:95) that directionality has to do with the translation direction of the texts that a corpus has. For instance, a corpus can comprise of texts originally written in L1 and their respective translations in L2. The direction of the translation functions is just one direction, so in this case, they are called unidirectional. If a corpus is made up of text originally written in L1 and their translations in L2 plus original in L2 and their translation in LI, this is called bidirectional. Multidirectional corpora are also possible especially, when more than two languages are involved and their translation direction is not centered on L1 but on the interaction among all the languages constituting the corpus.

CHAPTER TWO

LITERATURE REVIEW

2.0 Preamble

This section reviews the literature relating to the MT and language. Some of the issues considered are language in translation, computer learning processes and Statistical Machine Translation (SMT) and its models

2.1 Language in Translation

> I speak to you in your language, and it is in mine that I understand you (Edouard Gissant 1990 in Owolabi 2010:1).

The root of communication and the basic tenet of conceptual understanding are vested in the assertion above. In any instance where one of the interlocutors uses a word that is not represented in the language, idea, view, perception or imagination of others; there is bound to be a misrepresentation, misunderstanding or lack of communication, depending on the degree of vagueness of such word(s) used in the discussion. This is often the situation especially in a condition where the interlocutors are not using their mother tongue or first language to discuss.

Consequently, whenever an idea, concept in the culture or nuance of a language is not well represented in the language of either interlocutors, communication is lost and understanding such an idea or concept becomes difficult. For instance, Osundare (1995) observes that it is problematic to represent the culture of a language in another language. He said:

> …because each culture has its own way of looking at the general world and since culture and language are intimately related, problems are bound to arise when attempts are made to articulate one culture in the language of another.

This he claims to be the ordeal of African writers that write in European languages. The two major creative-stylistic strategies devised by the African writers for tackling the problem to Osundare are: translation and transference. Translation is defined by Catford (1965:1) as an operation performed on languages: a process of substituting a text in one language for a text in another. Catford's definition can be summarised as:

Translation = SL => TL

where **SL** is source language and **TL** is target language.

Catford’s definition has been seriously criticised (Jakobson 1959, Nida and Taber 1969 / 1982 and Osundare 1995). Critics believe that Catford’s definition is limited to linguistic texts and there is more to translation than mere linguistic text equivalence. For example, Osundare (1995) opines that African literary writers that write in European languages are translating more than texts. To him, what they are doing is stylistic translation. He explains that ‘text’ is not always available in the indigenous languages and cultures which is awaiting an uneventful transposition into English language. Rather, in stylistic translation, the writer attempts to render in English the figures and tropes of L1[2], striving consciously, and oftentimes laboriously, to preserve their original flavor, the rhythms and cadences of their sentences, their idiomatic and proverbial authenticity and even their situational-dramatic occasions. As much as what African writers like Soyinka, Achebe and others are doing may not necessarily be considered as text translation as Catford’s definition suggest , to what extent do they achieve stylistic translation as claimed by Osundare (1995)? To corroborate the question, he (Osundare) refers to a Yorùbá song that says:

O gb’Oyìnbó títí
O lè fi Gẹ́ẹ̀sì súfèé
Jẹ́ ká rí ọ nísiyí
Bóyá o lè f’Oyìnbó kifá

You are so versed in English
You can even whistle in English
Now let’s see
Whether you can chant Ifa divination in English

With Osundare’s example above, translation whether textual or stylistic has its challenges in language because each language, with its uniqueness, represents features, concepts, cultures and ideas of its immediate environment. Hence concepts, objects and ideas missing in such cultural environment may not necessarily have representation in such a language. For example “snow” has different words among the Eskimos, Finish, Norwegians and Swedish. But among the Yorùbás there is no specific word for it. Though it is generally translated as *yìnyín* because that is the closest match to snow in the language but *yìnyín* is not snow. *Yìnyín* is equivalent to ice. Therefore, translating words or concepts like “snow” is better loaned than looking for direct equivalence in

[2] L1 refers to first language or mother tongue

the language (in this case, Yorùbá language) since the concept is not represented in the culture, and the immediate environment.

2.1.1 Language, Culture and Environment

Explaining the interrelatedness of language and culture, Isola in Oladipo and Adeleke (2010:1&2) is of the view that a language cannot be separated from its culture. This he explains in relation to the deplorable behaviuor and poor attitude of some Yorùbá educated elite towards their language. He explains further that:

> What is missing so far is the lack of emphasis on the role of language as the centre of the culture. Language is the hub of the wheel of culture while all other aspects like administrative, judicial, religious, educational and other system are the spokes. When a language dies, the culture dies.

Sapir (1912) links language and culture with the environment. He opines that:

> There is a strong tendency to ascribe many elements of human culture to the influence of the environment in which the sharers of that culture are placed, some even taking the extreme position of reducing practically all manifestations of human life and thought to environmental influences.

He categorises environmental influences into two: the physical and the social factors. To him, any description of human culture as due solely to the force of physical environment rests on a fallacy. Hence, the environment can act directly only on an individual, and in cases where a purely environmental influence is responsible, a communal trait is found, the common trait must be interpreted as a summation of distinct processes of environmental influences on individuals. However, a single individual can be truthfully said to be open to environmental influence. This being uncombined with the influence of another character is doubtful but at least possible. Therefore, the smallest environmental influence is either supported or transformed by social forces. On the other hand, the social forces may be looked upon, somewhat metaphorically, as parallel in their influence to those of heredity in so far as they are handed down from generation to generation. So, the physical environment is reflected in language only in so far it has been influenced by social factors.

He goes further to state that language may be influenced in three ways: subject matter or content, (that is vocabulary); phonetic system, (that is the system of sounds with which it operates in the building of words), and grammatical form (that is the formal processes and logical or psychological classifications made use of in speech).

He concludes that there seems to be no correlation between physical and social environment, and phonetic systems either in their general acoustic aspect or in regard to the distribution of particular phonetic elements. He says that:

> We seem, then, perhaps reluctantly, forced to admit that, apart from the reflection of environment in the vocabulary of a language, there is nothing in the language itself that can be shown to be directly associated with environment.

In other words, other two areas of environmental influence on language, speech sound and grammar, are jettisoned.

There are three main issues about Spair's (1912) position:

I. the influence of environment on language;
II. the ways language may be influenced and
III. the main thesis of his claims

I **The Influence of Environment on Language**

This influence which depends on physical and social factors is incorporated in interest; the term interest is however questionable. Though it is explained that the physical environment is inflected in language only in so far as it has been influenced by social factors; this too has to depend on *interest* of the member of the community before a linguistic symbol could be assigned to such an item or element. The question, therefore, is, how do the members of the community decide and conclude on common interest before a linguistic symbol is assigned to any item? Do they have to agree on words that are synonymous or antonymous in meaning? How does a child show interest in the passing down of languages since social factors are handed down from generation to generation? What becomes of a child or a new language learner who refuses to show interest in an item already named? Will that mean that the name of the item has to be changed for such an individual or will the person forget the name of such an item? With regards to this, Sapir did not explain the role of interest in the influence of the environment on language.

Language is complex, there might also be a degree of environmental influence on language but it is unexplainable that a group of people will not have interest in an item existing in their locality. Evidence of this is seen in the fact that there is no item whether of interest or not that such people do not have words for. In a situation where there is language or cultural contact, there will be the need to look for equivalences for such items in the language or loanwords from the language of the item or the neighboring language.

II The Ways Language may be influenced

Sapir (1912) itemizes three ways by which a language can be influenced vocabulary, speech sound and grammar. He concludes that the impact of the environmental influence on language is only felt on vocabulary and not necessarily on speech sound and grammar. The concern is how does a child acquire a speech and grammar of a language if the child is not in the environment where the language is spoken? Sapir explains that there seems to be an absolute lack of correlation between physical and social environment, and phonetic system, either in their general acoustic aspect or with regard to the distribution of particular elements. This he considers as an accidental character of a phonetic system. He goes further to explain that the fact that a phonetic system may be thought to have a quasi-mechanical growth, at no stage subject to conscious reflection and hence not likely in any way to be dependent on environmental conditions or, if so, only in a remotely indirect manner. But the fact remains that though Sapir considers the influence of the physical environment without much consideration of the social environment. The social environment provides the child with elements of the language for appropriate language acquisition. For example, the child gets the vocabulary, speech sound and grammar from the people in the immediate environment.

III The Main Thesis of His Claims:

The main thesis of Sapir (1912) is the complexity and rapid change in culture may not necessarily reflect on language. He points out that, cultural elements serve the immediate needs of the society and entering more clearly into consciousness will not only change more rapidly than those of language, but the form itself of culture, giving each element its relative significance, will be continually shaping itself anew. Linguistic elements on the other hand, while they may and do readily change in themselves, do not so easily lend themselves to regrouping owing to the subconscious character of grammatical classification. A grammatical system as such tends to persist indefinitely and therefore has the conservative tendency to make itself felt more profoundly in the groundwork of language than of culture. He maintains that the consequence of this is that the form of language will in the course of time cease to symbolise those of culture. Though he is of the position that the rapid change in culture will correspond but not equally to the rapid change in linguistic forms and contents. This is the direct opposite of the general view held with respect to the greater conservatism of language in civilised communities than among the primitive people. Then, what is the yardstick to measure civilised and primitive peoples' language if Sapir holds the view that he doubts whether many languages of primitive people have undergone a rapid modification in a corresponding period of time as has the English language?

To this researcher, grouping languages as civilised or primitive is a super imposition of a language over and above others. To Pinker (1995), in Fee (2003), assuming that there might be an existence of important language-based differences among cultures is considered colonialisation. Every language should and can express thoughts which proves the equality of all languages. Therefore, categorising languages into major, minor, civilised and primitive is basically for political and social reasons and not necessarily because a language is more important than the others. The school of thought that gives priority to this equality of language is that that views language as innate.

2.1.2 The Innateness of Language

The innateness of language which is championed by biolinguists is devoid of response to stimulus or environmental influence; they advocate that language is basically biological. They stress the fact that language grows like any other organs in humans and emphasize the fact that language is encoded in the DNA, which is why we are good at it, like being "good" at having two arms. Martinsson (2012) explains that the 'language organ' is what enables humans to learn the abstract theories of a language without even thinking actively about it. Since the Primary Language Data[3] (PLD) will not be a complete representation of a language and children will, after a number of years, learn the grammar of that language, there must be something preventing them from making incorrect generalizations based on the PLD. However, a language (from a biolinguistic point of view) is not something that the mind represents but is instead a property of the mind; it is not something that the brain keeps or codifies. Rather is part of the structure of the brain (José-Luis Mendívil-Giró 2009).

[3] The Primary Language Data is the data children are exposed to while they are learning their native language.

Mendívil-Giró (2009:9) shows the difference between two types of theoretical linguistics in the table below:

Chomskyan Linguistics	**Functional Linguistics**
Internism	Externalism
Rationalism	Empiricism
Formalism	Functionalism
Universalism	Relativism

Table1: A table showing the difference between two types of theoretical linguistics (adapted from Mendívil-Giró 2009)

Mendívil-Giró explains further that functional and cognitive linguistics depend on a functionalist conception of the mind whereby linguistic expressions and even languages are not objects of study in themselves but are instead a means of representing reality or of communicating thoughts. From this viewpoint, linguistic expressions convey propositions that are by definition external to the mind and consequently lack intrinsic structure. It is supposed that languages lack structure beyond what is necessary to fulfill certain cognitive or communicative functions.

On the language as an internal property of the mind, he explains that the main implication is that languages are not necessarily the outcome of external factors, nor are they necessarily the outcome of their adaptation to any communicative need; rather, languages owe their structure, to some degree at least, to constrictions internal to the structure of the mind. Moreover, from a biolinguistic point of view, a language is a person's faculty of language which significantly implies that there is no substantial difference between language and languages—or, at least, no more than there is a distinction between life and living beings. The distinction between language and languages is artificial but it may be useful in some contexts.

Jackendoff (2012) takes another dimension to the view of biolinguist in the explanation of human language. He queries the capacity of Narrow Language Faculty (FLN) or Universal Grammar (UG) which is unlearned capacities specific to linguistic modality against Broad Language Faculty (FLB) which includes FLN plus other mental machinery necessary for acquisition and use of language which serve other cognitive purposes. He explains that the Minimalist Program which reconstructs syntactic theory around Merge as the central computational operation, building syntactic structures recursively, plus the mappings from syntax to the "sensory-motor interface" – the auditory and motor systems in language perception and production respectively – and to the "conceptual-intentional interface" – the formation of thought and meaning, or "general intelligence".

He is of the opinion that redundancy is a characteristic of the brain which changes the desiderata for linguistics theory hence squeezing redundancy out of linguistics representations and rendering linguistic processing unnecessary. He further explains that Binary Merge is not rich enough to capture the varieties of recursion found elsewhere in mental structure. Also, a grammar stated in term of constraint satisfaction is preferable to one stated in terms of algorithmic generation and the brain's characteristic combinatorial possibilities which is Unification rather than Merge. He proposes that instead of Minimalist Program that super imposes Syntax well and above semantics and phonology which are regarded as the two interface systems of output representations, i.e. the semantic/LF component and phonetic/PF component. The Parallel Architecture which allows the three components to be independent of each other. Consider the illustrative scheme below:

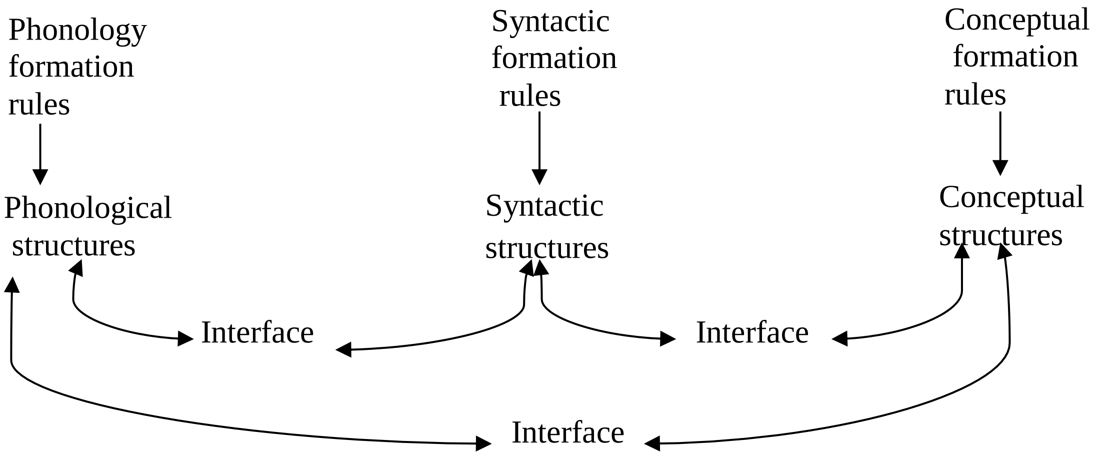


Fig. 2: The Parallel Architecture (adapted from Jackendoff 2012)

From the illustrative diagram labeled figure (1) above, it is observed that phonology, syntax, semantics and conceptual structure are independent generative systems linked by interfaces which are the license correlations between pieces of structure in two or more distinct domains.

He buttresses this further that words like *hello, upsy-daisy*, *gee whiz, feh* and *dammit* have phonological and semantical relevance, but do not participate in syntactic combinations. Hence, there is no reason for them to have syntactic features. Other interface between syntax and conceptual structure also include canonical principles for argument realization such as the thematic hierarchy; the interface between syntax and phonology which includes the principles that establish possible correspondence between syntactic structure and prosodic contours. In short, a well-formed sentence has well-formed structures in all three domains, and well-formed links between the domains.

Chomsky, in his Minimalist Approach to human language, does not superimpose syntax on semantics and phonology as Jackendoff (2012) claims; rather, it shows their interrelatedness. For example Cook and Newson (2007) clearly show this interrelatedness in their explanation of computational system of language from lexicon in the figure (2) below:

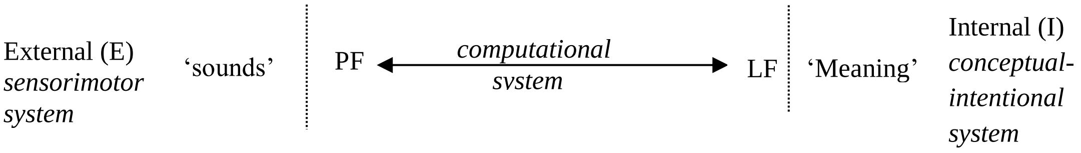


Figure 3: The Computational System (Adapted from Cook and Newson 2007)

The diagram above shows that computational system marries both PF and LF to generate well-formed sentences not necessarily superimposing it. In another example, *hello*, which Jackendoff (2012) explains, does not have a syntactic feature as used in sentence numbered (1) below:

1. She said "hello"

If *hello* enters the derivation, it will be merged with the verb. The verb *said* has grammatical features and it has to discharge it, ACC-case features which means that *hello* would be raised to spec VP for ACC-case valuation. Thus the un-interpretable features [Num,] [Per] and [Gen] on both the verb *said* and the DP object *hello* would be valued and deleted. Consequently, ACC-case feature is a by-product of valuation of Phi-features.

On participation in Merge, every lexical item is assumed to be fully specified in the lexicon. This means that every item entering a derivation has all the required features to participate on Merge operation and Merge is an economy operation which derivation has to be determined in order to reduce the computation burden.

2.2 Language Acquisition Theories and Machine Learning

The major two opposing views on language acquisition theories are: nature and nurture, behaviourism and nativism. Exponents of nurture/behaviourists/non-nativist/ empiricist believe that language acquisition is a set of learned habit, Nativist/Innatist are of the opinion that language is biological in nature (see Scholz and Pullum 2005; Clark 2008; Diessel 2008; Mendívil-Giró 2009; Fitch 2009 and 2011). The common phenomenon to the opposing view is the environment. While empiricists explain that environment is the teacher from which the child learns; the nativists believe that there is a need for the environment to nourish and trigger the innate ability. This unifying phenomenon that is the environment, can be regarded as data in machine learning (a branch of Artificial Intelligence). In the explanation of artificial intelligence when it comes to natural language processing, machine learning is seen as data driven; the more the data, the better the output. In an attempt to define Machine Learning (ML), Omary and Mtenzi (2010) opine that its definition must include these two critical elements: computer-based knowledge acquisition process and the knowledge source. Alpaydin (2004), as quoted in Omary and Mtenzi (2010), defines *Machine Learning* as the capability of the computer programme to acquire or develop new knowledge or skills from existing or non existing examples for the sake of optimising performance criterion. Apaydin's definition fulfils the two critical elements though computer based knowledge acquisition process is not clearly spelt out.

Machine learning (ML) concentrates on the theoretical foundations of learning from data (Solomatine and Ostfeld 2008). Domingos (2012) explains that ML which is also known as data mining or predictive analytics is an attractive alternative to manually constructing programmes. He emphasis that it spread rapidly across computer science and beyond in the last decade. He points out areas that ML is used as web search, spam filters, recommender systems, ad placement, credit scoring, fraud detection, stock trading, drug design and many other applications. He predicts from a report from the McKinsey Global Institute that ML will be the driver of the next big wave innovation. Smola and Wishwanathan (2008:4-5) link Machine Translation to ML. They explain that one of the applications of ML is automatic translation:

> An equally ill-defined problem is that of automatic translation of documents. At one extreme, we could aim at fully understanding a text before translating it using a curated set of rules crafted by a computational linguist well versed in the two languages we would like to translate...

> Instead, we could simply use examples of translated documents, such as the proceedings of the Canadian parliament or other multilingual entities (United Nations, European Union, Switzerland) to learn how to translate between the two languages. In other words, we could use examples of translations to learn how to translate. This machine learning approach proved quite successful.

By implication, the data which the machine learns from is the environment that the language acquisition theorists are arguing about which in actual sense is their unifying point which in conclusion nurture language acquisition for humans and triggers learning for machine. As mentioned in chapter one, the learning could be supervised or unsupervised. (Pantic Solomatine and Ostfeld 2008, and Taiwo 2010)

2.2.1 Supervised Machine Learning

Supervised learning comprises of algorithms that reason from externally supplied instances to produce general hypothesis which then make predictions about future (Omary and Mtenzi 2010) . In other words, it entails learning a mapping between a set of *input* variables *X* and an *output* variable *Y* and applying this mapping to predict the outputs for unseen data (see Cunningham, Cord, and Delany 2008). This means that in supervised learning there is a presence of the outcome variable to guide the learning process. Cunningham, Cord, and Delany (2008) link supervised learning to statistics, they explain that in the supervised learning paradigm, the goal is to infer a function $f : X \rightarrow Y$, the classifier, from a sample data or training set A_n composed of pairs of (input, output) points, $\mathbf{x}_i$ belonging to some feature set X , and $y_i \in Y$:

$$A_n = ((\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)) \in (X \times Y)^n.$$

Typically $X \subset IRd$, and $y_i \in IR$ for regression problems and y_i is discrete for classification problems.

2.2.2 Unsupervised Learning

Unlike supervised learning, unsupervised learning builds models from data without predefined classes or examples. This means no “supervisor” is available and learning must rely on guidance obtained heuristically by the system examining different sample data or the environment. The output states are defined implicitly by the specific learning algorithm used and built in constraints (Omary and Mtenzi 2010). This means that machine models a set of inputs without the availability of labelled examples to classify or map out a vector into classes.

There are other types of ML algorithms like semi-supervised learning, reinforcement learning, Transduction and Learning to learn (see Taiwo2010). Lopez (2008) also links LM to MT. He stresses that SMT is an approach to MT that is characterized by the use of machine learning methods. In other words, the machine learning methods includes the aforementioned. In the next section, we shall explore Statistical Machine Translation and its types.

2.3 Statistical Machine Translation (SMT)

The current trend in the MT research discourse favours SMT because there are now the availability of corpora as well as cheaper computing power with the introduction of statistical models which bring about SMT flexibility for adaptation to any language pair, coupled with its being less time consuming and having reduced cumbersome rules (see Mylonakis 2012, Kohen 2010 and Lopez 2008). SMT is an example of a Corpus-Based MT model; it treats the translation of natural language as a machine learning problem. This means that learning algorithm is applied to a large body of previously translated text known variously as a *parallel corpus, parallel text, bitext*, or *multitext*. The machine is then able to translate previously unseen sentences (Lopez, 2008).

The system examines many samples of human-produced translation and the SMT algorithms automatically learn how to translate. In less than two decades, SMT has come to dominate academic MT research, and has gained a share of the commercial MT market. Progress is rapid, and the state of the art is a moving target. However, as the field has matured, some common themes have emerged (see Lopez 2008; Tripathi and Sarlchel 2010; Koehn 2010; Mylonakis 2012). Mylonakis (2012:18) explains the basic process of SMT. He states that given a source language sentence *f*, the fundamental problem in MT is to produce its target language translation *e* by means of a computer program. Output *e* must both sufficiently convey the meaning of the original sentence *f*, as well as enjoy target language fluency. He emphasizes that SMT aims to achieve this through the application of statistical models. By introducing a probability distribution p(e|f), assigning to every target sentence *e* a probability of being the translation of source input *f*, an SMT system outputs the target sentence ê with the highest conditional probability:

$$\hat{e} = \arg\max_{e} p(e/f)$$

To achieve this, Lopez (2008) gives four important ideas:

1. We must describe the series of steps that transform a source sentence into a target sentence. We can think of this as creating a story about how a human translator might perform this task. This story is called a *translational equivalence model,* or more simply a *model*. All of the translational equivalence models that we will consider derive from concepts from automata and language theory
2. Next, we want to enable our model to make good choices when faced with a decision to resolve some ambiguity. We need to develop a *parameterization* of the model that will enable us to assign a score to every possible source and target sentence pair that our model might consider
3. The parameterization defines statistics set called *parameters* used to score the model but we need to associate values to these parameters. This is called *parameter estimation*, and it is based on machine learning methods
4. Finally, when we are presented with input sentence, we must search for the highest scoring translation according to our model. This is called *decoding.*

Mylonakis (2012) however reduces Lopez's ideas to three steps:

1. This involves designing the model p(e|f). The fundamental questions to be asked are: what kind of translation phenomena does the model capture and how does it capture them? What are the parameters? And which latent variables are assumed (because model design plays a crucial role in SMT as it defines the rules of the game: what needs to be learnt from the training corpora and later applied to actually translate, according to the modellers view of translation). After the model is set, the two other steps below must be considered.
2. We need to train it, select the model instance which is best according to some learning objectives by employing training data possibly coupled with prior knowledge. This entails the usage of a statistical estimator.
3. Then, the final step, which is decoding, employs the trained model estimate to actually translate by selecting for every input f the translation ê according to the equation given above.

You will observe that Mylonakis (2012) collapses (1) and (2) of Lopez (2008) as (1) in his idea showing basically that both of them are saying the same thing. Hearne and Way (2011) see Lopez’s (2008) explanations as well-defined decision problem which can be scored in two ways: the noisy-channel model and the log-linear model. According to them, the noisy-channel model is traditionally used in the literature while the log-linear model can be instantiated to express precisely the same computation as the noisy-channel one; it is more flexible and has come into widespread use in recent years. We shall then look at the noisy-channel and the log-linear models below.

2.3.1 Noisy-Channel Model

$$\mathit{Translation} = \mathit{argmax}_T\, P(f|e).\, P(e)$$

According to Koehn (2010) Noisy-channel model comprises of two component scores P(f|e) and P(e) which are to be concatenated. The first component is, P(f|e), that is, the likelihood that the source sentence f and the target sentence translation e are translationally equivalent, meaning that the meaning of both f and e are captured as expressed in the text (i.e translation adequacy). The component is generally referred to as the translation model. The second component, P(e), implies the likelihood that the target sentence translation e is actually a valid sentence in the target language (fluency) and is generally referred to as the language model. Mylonakis (2012) however states that early SMT works, such as the IBM models applied the Noisy Channel paradigm in a relatively literal fashion. He opines that translation adequacy and fluency can in practice hardly be considered separate. This is that malformed target outputs cannot appropriately convey the meaning of the source sentence; an adequate translation would probably be expected to also be relatively well-formed

2.3.2 Log-Linear Model

$$\mathit{Translation} = \mathit{argmax}_T \sum_{m=1} \lambda_m \cdot h_m\, (e,f)$$

The log-linear model is more general than the noisy-channel model in that it expresses scoring in terms of aggregation of an unlimited number of feature scores. First, a technical note: the log-linear model uses log probabilities rather than regular ones. They are converted to log probabilities simply by applying the log function – and log probabilities are added rather than multiplied. Thus, generally log(X·Y) = log(X) + log(Y) and, more specifically with respect to the noisy-channel model, log(P(f|e)·P(e)) = log(P(f|e)) + log(P(e)).

Log-linear model equation $\sum_{m=1}^{M} \lambda_m . h_m$(e,f) has a set of log feature scores to be added together.

$\sum_{m=1}^{M}$ notation indicates that there are a total of M features to be scored and that their individual scores are to be added up (the sigma, Σ). These individual scores are to be computed by multiplying two feature-specific values, ʎ$_{m}$ and h$_{m}$(e,f), where ʎ$_{m}$ is simply a weight indicating the importance of that feature relative to the other features, and h$_{m}$ (e,f) is the log probability assigned to the source–candidate pair by that feature. A minimum of two features are usually used: the translation model and the language model features, just like the noisy-channel mode.

2.3.3 Generative and Discriminative Models

Brown et al (1993), in Mylonakis (2012), explains generative translation model as a model which captures the stochastic joint generation of source and target sentence pairs. They can also straightforwardly be employed to select the translation *e* with the highest probability given *f*, as with *f* fixed:

$$\hat{e} = \text{arg max } p(\text{e/f}) = \underset{\text{e}}{\text{arg}} \text{ max } p(\underset{\text{e}}{\text{e}}\text{/f})$$

These models are usually based on a generative process tracking the steps to emit the tuple [e,f]. For example, we might begin by considering the generation of corresponding source and target word-pairs following the word order of the source language, and subsequently reordering the target language words to form the target sentence. Each of the generative steps is modelled by a separate distribution conditioned on the previous steps, often under independent assumptions which simplify the modelling effort. Some conditional translation models p(e|f) are formulated in a similar fashion, emitting *e* from *f* under a generative process. Generative models require extensive efforts to consider all the steps and transformations that take place during translation, as well as to introduce independence assumptions taking into account the available training data (e.g. to avoid overfitting) or computational limitations etc.

In contrast, discriminative modeling directly models the conditional distribution p(e|f), instead of putting effort towards formulating a fully generative process emitting samples (e,f). For MT, this typically happens through employing feature functions ϕ_i (e,f), each assigning a non-negative score as well as examining the two sentences from a different perspective, word or phrase correspondence, output fluency (frequently the

LM score), target word reordering and others. The modeller does not need to consider a coherent generative story but only what kind of features could be useful in discriminating between strong and weak translations. These scores are weighted together log-linearly with weights λ_i and normalised to obtain the conditional translation model (Och and Ney, 2004)

You will observe that *Translation = arg max* is common to all the models indicating that the translation of the target sentence *T* is the maximum or best (*argmax*) score of the translation. This is the target of both models. One of the most important components of SMT to achieve *argmax* is the language model.

2.2.4 Language Model

"One essential component of any SMT system is the **language model**, which measures how likely it is that a sequence of words would be uttered by a language speaker. It is easy to see the benefits of such a model. Obviously, we want a MT system not only to produce output words that are true to the original in meaning, but also to string them together in fluent sentences" (see Koehn 2010:181).

Besides the output string which must reflect the fluency of the speakers of the language e.g it is likely that a Yorùbá person will say (1&2) rather than (3&4)

1. Mo fẹ́ jẹun
   1sg want eat
   I want to eat

2. Mo fẹ́ Ìyábọ̀
   1sg love Ìyábọ̀
   I love/marry/want Ìyábọ̀

3. *Mo jẹun fẹ́
   1sg eat want

4. *Ìyábọ̀ Mo fẹ́
   Ìyábọ̀ 1sg love
   Ìyábọ̀ Mo fẹ́

It also helps to support difficult decision about word order and translating words, for example consider (5) and (6) below:

5. Mo fẹ́ lọ ilé
   I want go house
   I want to go home/house

6. Mo fẹ́ lọ ibùgbé
   I want go dwelling place
   I want to go to the dwelling place

You will observe that a Yorùbá language speaker will often say (5) than (6) in a conversation where his intension of going home is to be expressed though both (5) and (6) refer to the same place. Hence, a Probabilistic Language Model *PLM* should prefer the correct word order to the incorrect word order (that is, 1 and 2) and assign a higher probability to (5) than (6).

2.3.5 N-gram

The leading approach to language modeling is **n-gram** language modeling (Koehn 2010:182). N-gram is a statistical tool that finds the occurrence of a string (word) from the large corpus. In other words, it studies how likely words are to follow each other. For example; given a string of Yorùbá words W= w_1, w_2, w_3…; wn we need to find *p(W)*. *p*(*W*). However, the probability that if a sequence of words is picked at random it turns out to be *W*. To compute *p*(*W*) the chain rule presented below will be needed:

$$\boldsymbol{p(w_1,w_2w_3 \ldots;wn) = p(w_1)\ p(w_2|w_1)\ p(w_3|w_1,w_2) \ldots p(wn|w_1,w_2, \ldots wn\text{-}1)}$$

One of the chain rule used is the **Markov chain.** Koehn (2010:184) explains that the **Markov assumption** states that only a limited number of previous words affect the probability of the next word. It is technically wrong, and it is not too hard to come up with counter examples that demonstrate that a longer history is needed. However, limited data restrict the collection of reliable statistics to short histories. He stresses that typically, the number of words in the history is based on how much training data is available. More training data allows for longer histories. He notes that most commonly, **trigram (**which consider a two-word history to predict the third word language models) are used. This requires the collection of statistics over sequences of three words, so called 3-grams (trigrams). Language models may also be estimated over 2-grams (**bigrams**), single words (**unigrams**), or any other order of n-gram. The assumption of 3-gram however, is that words only relate with each other in respect to threes words around it. This is not so true with natural language. Consider (7) below:

7. She reported the incident to her husband.

Our knowledge of the English language gives us the awareness without much explanation that *she* and *her* are co-referential and they must agree in gender and number. Otherwise, the sentence will be ungrammatical. Looking at their position in the sentence, we will need 5-gram to show the relationship of *she* and *her* as found in (7) above which means we will need a very large memory to perform very small functions.

2.4 Models of Statistical Machine Translation

Models of SMT can be seen from two dimensions: the alignment models and the translation models (see Lopez 2008 and Brunning 2010). We shall focus on the translation model now and the alignment model in the next section. The translation model has much to do with the translation approach adopted for the SMT; the followings are the widely discussed translation models in the literature: Word-Based Model, Phrase-Based Model, Syntax-Based Model, and Synchronous Context Free Grammar Model.

2.4.1 Word Based Translation Model

This is a translation model that is based on lexical translation, that is, translation of words in isolation (Koehn 2010:81). It is the first generation of SMT models introduced by IBM models (1-4) (Lopez 2008, Brunning 2010, and Koehn 2010). In other words, SMT uses probability and the highest statistics to translate; a word may have more than one equivalent translation if we have to check from a bilingual dictionary. For example, *A Dictionary of Yoruba Language* published by the University Press Limited in 1950 has the followings as the equivalence of *ilé*: house, home, mansion, and dwelling. Koehn (2010:82) explains that the concern of SMT is, what are the possible translations and how often do they occur? The question will lead to estimating the probability distribution of each of the translations of *ilé* in a given data. Adapting an example from Koehn (2010), we could have a hypothetical distribution table of the translations in the table below:

Table 1:

Translation of *ilé*	**Count**
House	6000
Home	2000
Mansion	1500
Dwelling	500

Table 2: A diagram showing equivalences of ilé in English and their distribution paterns

We may want to estimate a lexical translation[4] probability distribution from these counts. This will assist to answer a question that may arise when we have to translate a new Yoruba text: what is the most likely translation for a foreign word like *ilé*? In other words, we want to find a function:

$$pf: e \rightarrow pf(e)$$

that is, given a foreign word *f* (here *ilé*), return a probability, for each choice of English translation *e*. The function should return high value if an English candidate word *e* is a common translation. It returns a low value if an English candidate word *e* is a rare translation. It returns 0 if the English translation *e* is impossible. The definition of probability distribution requires the function *pf* to have two properties:

$$\sum_{e} pf\,(e) = 1$$

$$\forall e\ :0 \leq pf\,(e) \leq 1$$

Deriving the probability distribution from Table 1 above, we could use the ratio of counts. We have 10,000 occurrence of word *ilé* in our hypothetical text. In 6000 instances, it is translated as *house*. Dividing these two numbers, the result is 0.6, so we could set the $p_{ilé}$(house) = 0.6. If this is done for the rest of the three choices, we could have:

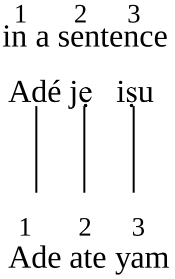
$$Pf(e) \;=\begin{cases} 0.6 & \text{if } e = \text{house} \\ 0.2 & \text{if } e = \text{home} \\ 0.15 & \text{if } e = \text{mansion} \\ 0.05 & \text{if } e = \text{dwelling} \end{cases}$$

Koehn (2010) explains that this method of obtaining a probability distribution from the data is not only very intuitive; it also has a strong theoretical motivation. Other ways of building a model for a given data is **Maximum Likelihood Estimation** whereby the probability mass for unseen events is reserved.

[4] Lexical Translation is used synonymously as Word Translation in the literatures

IBM model 1 is associated with alignment (see Lopez 2008 and Koehn 2010). It is a generative model[5] for sentence translation based solely on lexical translation probability distribution. It allows the definition of a model that generates a number of different translations for a sentence, each with a different probability. Word Alignment, however, is a microcosm of translation (Lopez 2008:25). Lopez (2008) explains that the word alignment can be viewed as a substitute for decoding, since it is more constrained – because, in word alignment, a correspondence is found between sequences, whereas in decoding we will be required to find both the correspondence and the target sequence. Summarily, a word alignment task is to discover the word-to-word correspondence in a sentence pair (e^I,f^J). For example:

8.



You will note that translation is done from Yorùbá to English and the alignment maps English word position to Yorùbá word position. This function provides a mapping of this nature:

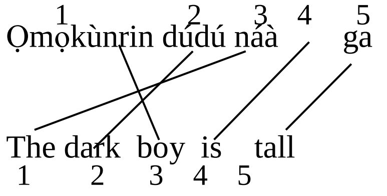
$$a: \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3\}$$

An alignment can be formalized with alignment function *a* (Koehn 2010:84). The function maps each English output word at position *i* to a Yorùbá input word at position *j*:

$$a: j \rightarrow i$$

We may not be able to map all the time like the example (8) above which will warrant reordering rule translation as shown in the example (9) below:

8.



$$a: \{1 \rightarrow 3, 2 \rightarrow 2, 3 \rightarrow 1, 0 \rightarrow 4, 5 \rightarrow 5\}$$

[5] breaking up the process of generating the data into smaller steps, modeling the smaller steps with probability distributions, and combining the steps into a coherent story – is called Generative modeling (Koehn 2010:86).

You will observe that mapping of (9) is not like (8) above. The word order of (9) is slightly changed. This will warrant a reordering rule and the need for Null token that '*is'* is mapped with, in the source language. Hence, a function word, '*is'*, does not have a clear equivalent in Yorùbá, so it is marked with a Null token. Therefore, an alignment model allows for dropping, adding and duplication of words during translation (see Koehn 2010: 85-86).

To factor in alignment model to translation probability *p(e|f)*, translation probability is to be defined for a foreign word **f** = $(f_1,\ldots,f_{lf})$ of length of *lf* to an English sentence *e* = $(e_1, \ldots,e_{le})$ of length *le* with an alignment of each English word e_j to a foreign word f_i according to alignment function $a : j \rightarrow I$ as follows:

$$p(\mathbf{e},\mathrm{a}|\mathrm{f}) = \frac{\epsilon}{(lf+1)^{le}} \prod_{j=1}^{l_e} t(e_j|f_a(j))$$

Koehn (2010:87) explains that the most important part of the lexical translation probabilities for all l_e is to generate output words e_j. The fraction before the product is necessary for normalization. Therefore, since the special NULL token is included, there will be $l_f + 1$ input words. Hence, there are $(l_f + 1)^{le}$ different alignments that map $l_f + 1$ input words into *le* output words. The parameter ε is a normalisation constant, so that $p(\mathbf{e},a|\mathbf{f})$ is a proper probability distribution, meaning that the probabilities of all possible English translations **e** and alignments *a* sum up to one:

$$\Sigma_{\mathrm{e,a}}\, p(\mathrm{e,a|f}) = 1$$

Applying this to example (8) above and its hypothetical translation probability table below:

adé

E	*t*(*e*|*f*)
Ade	0.9

jẹ

e	*t*(*e*|*f*)
yam	0.8

iṣu

e	*t*(*e*|*f*)
ate	0.7

three words are translated and therefore three lexical translation probabilities must be factored in:

$$p(e,a|f) = \frac{}{\epsilon} \text{ x } t(\text{adé}|\text{ade}) \text{ x } t(\text{jẹ}|\text{ate}) \text{ x } t(\text{iṣu}|\text{yam})$$

$$= \frac{\epsilon}{4^3} \text{ x } 0.9 \text{ x } 0.7 \text{ x } 0.8$$

$$= 0.007875\epsilon$$

So, the probability of translating the Yorùbá sentence *Adé jẹ iṣu* is 0.0079є using IBM Model 1 which is designed specifically for Word Based Statistical Machine Translation. It should be noted that there is a little difference in the denominator of Koehn (2010:87) and a version of the book on smtmt.org. While Koehn (2010) used 5^4, the online version used 4^3 for same sentence translation though the outcome is not much of difference.

2.4.2 Phrase Based Translation Model

The best performing Statistical Machine Translation systems are based on phrase-based models that translate small word sequences at a time (Koehn 2010:127). In other words, unlike the Word-based Translation Model that translates each word as an atomic unit, the Phrase-based Translation Model translates phrases that is contiguous sequences (Lopez 2008:8), sequences of consecutive words (Brunning 2010:13) or better still, contiguous multiword sequence (Koehn 2010:148) as atomic unit. It should be noted and emphasised that the contiguous sequence are not grammatically or linguistically motivated (Koehn and Knight 2003, Lopez 2008, Brunning 2010, and Koehn 2010,) as will be shown shortly in the examples below.

Zens and Nye (2004) opine that Word-based Translation Model loses contextual information. They note that lexicon probabilities are based only on single words and for many words; translation depends heavily on the surrounding words. Hence, Word-based Translation Model is not capable of addressing disambiguation through the language model. However, they suggest tne Phrase-based Translation Model which incorporate context into the translation by learning translations for a whole phrase instead of single words. They conclude that the basic idea of phrase-based translation is to segment a given source sentence into phrases, then translate each phrase and compose the target sentence from these phrase translations. Koehn (2010:127-129) enumerates the advantages of Phrase-based Model over Word-based Model as follows:

- Words may not be the best candidates for the smallest units for translation. Sometimes, one word in a foreign language translates into two English words, or vice versa.
- Translating word groups instead of single words helps to resolve translation ambiguities.
- If we have large training corpora, we can learn longer and longer useful phrases, sometimes even memorise the translation of entire sentences.

- The model is conceptually much simpler since such complex notions of fertility, insertion and deletion of the word-based model are done away with.

Blunsom (2009) is of the opinion that phrase based approach to SMT improves the modelling of multi-word translation units, increase contextual input; permits idioms and non-compositional phrases as well as eases search and reliance on the language.

To Lopez (2008:8) Phrase-based Model translation process takes three steps:

(1) the sentence is first split into phrases;

(2) each phrase is translated.

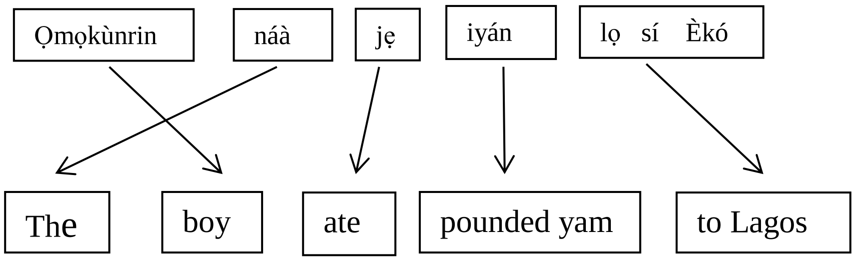
(3) the translated phrases are permuted into their final order.

The permutation problem and its solutions are identical to those in word-based translation. Let us consider example (10) below:

10. Ọmọkùnrin náà jẹ iyán lọ sí Èkó
child man the eat pounded yam go prep Lagos
The boy ate pounded-yam to Lagos

(10) is broken into phrases though not grammatically motivated as (11) below and are further translated based on the phrases which may also warrant reordering as Lopez (2008:8) explained.

11.



You will observe that (11) is not grammatically motivated. Koehn (2010:128) states that, the basic motivation is the context which is necessitated by the phrase translation table. The power of phrase-based translation rests on a good phrase translation table (Koehn 2010: 130). For example, the phrase table for iyán could be:

Translation	Probability
Pounded yam	0.7
, pounded yam	0.15
, pounded yam,	0.03
Yam that is pounded	0.05

In other words, the probability of translating the Yoruba word, *iyán*, to English *pounded yam* is 0.7 or in mathematical notation p(iyán | pounded yam) = 0.7. Note that these translation probabilities are in the inverse order due to the noisy channel model (see http://www.statmt.org/moses/?n=Moses.Tutorial).

According to Koehn (2010:129, Bayes rule is to be first applied to invert the translation direction and integrate a language model p_{LM}. Therefore, the best English translation $\mathbf{e}_{\mathrm{best}}$ for a Yoruba input sentence **y** is defined as

$$\begin{aligned} \mathrm{e}_{\mathrm{best}} &= \mathrm{argmax}_{\mathrm{e}}\, p(\mathrm{e}|\mathrm{y}) \\ &= \mathrm{argmax}_{\mathrm{e}}\, p(\mathrm{e}|\mathrm{y})\, p_{\mathrm{LM}}\, (\mathrm{e}) \end{aligned}$$

This is not quite different from the word based model though the phrase based model p(y|e) is further decomposed into

$$p(y^{-I}{}_{1}|e^{-I}{}_{1}) = \prod_{\mathrm{i=1}}^{1} \phi\, (\bar{y}_i|\bar{e}_i)\; d(\mathrm{start}_i - \mathrm{end}_{\;i\text{-}1} - 1)$$

Yoruba sentence y is broken up into I phrase $\bar{y}_i$ whereby each Yoruba sentence $\bar{y}_i$ is translated into English phrase $\bar{e}_i$ in the noisy channel so that the phrase translation probability $\phi(\bar{y}_i|\bar{e}_i)$ is modeled as a translation from English to Yoruba. He explains further that reordering is handled by thr distance based reordering model. The distance-based reordering model consider reordering relative to the previous phrases. So, start*i* is defined as the position of the first word of the Yoruba input phrase that translates to the *i*th English phrase, and end*i* as the position of the last word of that Yoruba phrase. Reordering distance is computed as $\mathrm{start}_i - \mathrm{end}_{i-1} - 1$. The reordering distance is the number of words skipped (either forward or backward) when taking foreign words out of sequence. If two phrases are translated in sequence, then $\mathrm{start}_i = \mathrm{end}_{i-1}+1$; i.e., the position of the first word of the phrase *i* is the same as the position of the last word of the previous phrase plus one. In this case, a reordering cost of $d(0)$ is applied.

Koehn (2010:130) defines *d* in line with the exponential decaying cost function $d(\mathrm{x}) = \alpha^{|\mathrm{x}|}$ instead of estimating reordering probabilities from data. He emphasises that with appropriate value of parameter $\alpha\ \epsilon[0,1]$ so that d is a proper probability distribution. This implies that movements over large distance are more expensive that shorter movements or no movement at all. He concludes that only the phrase translation table is learnt from data, reordering is handled by a predefined model.

2.4.2.1 Phrase Translation Table

Hardmeier (2010) citing Koehn et al (2003 & 2007), explains that the Phrase-based Statistical Machine Translation uses translation models in the form of phrase tables in which phrase pairs consisting of source language (SL) and target language (TL) word sequence, *s*, and *t*, are associated with a number of scores corresponding to different models of translation probabilities between *s* and *t*. Therefore, candidate phrase pairs are usually extracted from a parallel corpus with automatically generated word alignments. The forward and reverse conditional phrase translation probabilities p(s|t) and p(t|s) are then estimated by the relative frequency of SL phrase in alignment with a given TL phrase and vice versa. To overcome the unreliability of the estimates for low-frequency phrases, phrase tables usually include maximum likelihood scores for both p(s|t) and p(t|s) as well as two additional lexical weight scores based on the word alignment probabilities of individual component words of the score and the target phrases.

Hardmeier (2010) points out that *moses* as a toolkit for SMT has a tool called phrase-extract to extract phrase pairs from a word-aligned corpus and compute phrase translation probabilities and lexical weights. However, Koehn (2010) states that the power of the phrase-based translation rests on a good phrase translation table. Phrases are mapped one-to-one on a phrase translation table. He gives detailed explanation of a method of acquiring such a table. A word alignment has to be created between the sentence pair of the parallel corpus and then extract phrases pairs that are consistent with this word alignment. For example *lọ sí èkó* to mean *to Lagos* considering example (11) above which is repeated below for convenience as example (12).

12 Ọkunrin náà jẹ iyán lọ sí Èkó
Man det eat pounded yam go prep Lagos
'The man ate pounded yam before to Lagos'

Zens and Ney (2004) explain how they extract their phrase as follows: they first train statistical alignment models using GIZA++ and then compute the Viterbi word alignment of the training corpus. They did this for both translation directions. Then they take the union of both alignments to obtain a symmetrised word alignment matrix. It is said that this alignment matrix is the starting point for phrase extraction. They give the criterion which defines the set of bilingual phrases of sentence pair as $(f^{j}; e^{I})$ and the alignment matrix as $A \subseteq J\ x\ I$ which are used in the translation system.

Phrase based translation $(f^{j}, e^{I}, A) =$ $(f_{i_1}^{j2}, e_{j_1}^{i2})$: $\{$

$$\forall (j,i) \in A: j_1 \leq j \leq j_2 \leftrightarrow i_1 \leq i \leq i_2$$
$$\wedge \exists (j,i) \in A: j_1 \leq j \leq j_2 \wedge i_1 \leq i \leq i_2 \}$$

They emphasised that this criterion is identical to the alignment template criterion described in Och et al (1999) which indicate that the phrases are considered to be translations of each other and the words are aligned only within the phrase pair and not to words outside; and the phrases have to be contiguous.

Koehn (2010:136) gives a description of the size of the phrase table. He states that, for large parallel corpora of millions of sentences, the extracted phrase translation tables easily require several gigabytes of memory which may be too much to fit into the working memory of a machine. This causes problems for estimating phrase translation probabilities and the use of these tables to translate new sentences. He maintains that not all phrase pairs have to be loaded into the memory for the estimation of the phrase translation probabilities. He further said that it is possible to efficiently estimate the probability distribution by storing and sorting the extracted phrases on disk. Consequently, only a small fraction of the translation table for the translation of a single sentence is needed and may be loaded on demand.

2.4.2.2 Zens and Ney Translation Model

Zens$_{1}$ and$_{1}$ Ney (2004) explain the description of the translation model they used. They state that, they had to introduce a hidden variable S which is a segmentation of the sentence pair $(f_1^{j}\ ;\ e_1^{I})$ into K phrase $(\tilde{f}k_1; \tilde{e}k_1)$. They also use one to one phrase alignment; one source phrase is translated by exactly one target phrase as seen below:

$$Pr(f_1^{j} | e_1^{I}) = \sum_{s} Pr(f_1^{J}, S | e_1^{I})$$

$$= \sum_{s} Pr(S | e_1^{I}) - Pr(f_1^{J} | S, e_1^{I})$$

$$\approx \max Pr(S|\, e^I_1) - Pr(\tilde{f}^k_S \mid \tilde{e}^k_1)$$

By implication, only translations that are monotone are allowed at the phrase level. Then phrase $\tilde{f}_1$ is produced by $\tilde{e}_1$, and phrase $\tilde{f}_2$ is produced by $\tilde{e}_2$ etc. They (Zens and Ney) further explain that the re-ordering was learned within the phrases during training. They also indicate that, there is no constraint on the re-ordering with the phrases.

$$\begin{aligned} Pr(\tilde{f}^k_1; \tilde{e}^k_1) &= \prod_{k=1}^{K} Pr(\tilde{f}_k|\tilde{f}^{k-1}_1, \tilde{e}^k_1) \\ &= \prod_{k=1}^{K} p(\tilde{f}_k|\tilde{e}_k) \end{aligned}$$

They assume a zero-order model at the phrase level and estimate the translation probabilities

$p(\tilde{f}\,|\,\tilde{e})$ via relative frequencies:

$$p(\tilde{f}\,|\,\tilde{e}) = \frac{N\,(\tilde{f},\tilde{e})}{\Sigma_{\tilde{f}'}\, N\,(\tilde{f},\tilde{e})}$$

where $N\,(\tilde{f},\tilde{e})$ denotes the count of the event that $\tilde{f}$ has been seen as a translation of ẽ. They emphasis that if one occurrence of $\tilde{e}$ has $N > 1$ possible translation, each of them contributes to $N\,(\tilde{f},\tilde{e})$ with *1/N* and the counts are calculated from the training corpus. Zens and Ney (2004) used bigram language model, assuming Bayes decision rule and they are able to obtain the following search criterion:

$$\begin{aligned} \hat{e}^I_1 &= \text{argmax} \left\{ Pr\,(e^I_1).\ Pr\,(f^J_1 \mid e^I) \right\} \\ &= \underset{e^I_1}{\text{argmax}} \left\{ \prod_{i=1}^{I} p\,(e_i \mid e_{i-1}) \right. \\ &\qquad \left. .\ \max_S p(S|e^I)\ .\ \prod_{k=1}^{K} p(\tilde{f}_k|\tilde{e}_k) \right\} \end{aligned}$$

$$\approx \ \text{argmax} \left\{ \prod_{i=1}^{I} p(e_i\,|e_{i-1}) \prod_{k=1}^{K} p(\tilde{f}_k|\tilde{e}_k)_{e^I_1,\, S} \right\}$$

For refinements, they describe two simple heuristics: the word penalty feature and the phrase penalty feature. The model scaling factors were optimised with respect to the mWER on the development corpus. They describe an efficient monotone search algorithm which is with the phrase level and it has the translation speed of more than 1000 words per second for Verbmobil task and for the Xerox task. For the Canadian Hansards task, the translation of sentence length 30 takes only 1.5 seconds. Therefore, there are no constraints on the re-ordering, and by implication, the translation process should require only local reordering. They reveal that German-English Verbmobil task outline the limitations of the monotone search. In that the free word order in German as well as verb group seems to be difficult to translate. They suggest ignoring the word order but focusing on looking at the mPER only, then the monotone search will be competitive with the best performing system. They further suggest an investigation on the usefulness of additional models. These include modeling the segmentation probability as well as slightly relaxing the monotonicity constraint in a way that will allow an efficient search of high interest. And in line with the IBM reordering constraints of the single-word based models, it will be better if a phrase could be skipped and translated later.

2.4.3 Syntax Based SMT

This model is based on translating syntax models instead of translating single words or strings of words. If the input and output languages have different syntactic structures, sequence models (word-based or phrase-based) have difficulties with the increased amount of reordering during translation, especially long-range reordering. We may want to define reordering rules based on syntactic annotations (part-of-speech tags or syntactic trees) which restructure the input sentences into the order of the output sentences. These rules may be devised manually or learned automatically from word-aligned sentence pairs annotated with syntactic markups. Linguistic annotations may also be exploited in a re-scoring approach. By generating n-best lists of candidate translations, we annotate these with additional linguistic markups, which allow preference to be given to translations that show more grammatical coherence with the input and more grammatical agreement within itself. Consideration is also given to overall syntactic parse probability (Koehn 2010: 27).

Syntax-based models have been effective in capturing the long-range reordering between language pairs with very different word orders like Japanese-English (Lee, Zhao and Luo 2010). In line with the aforementioned, Ahmed and Hanneman (2005) explain that the output of phrase-based models fails to capture long-range movement at a deeper level like the modifier movement between English and French. To help remedy this problem, and produce fluent outputs, syntax-based models aim at modeling the deeper level of structures at the two sides of the noisy channel. The reason behind this is that the statistical methods do not employ enough linguistic-theory to produce a grammatically coherent output (Och et al. 2003). This is because the methods incorporate little or no explicit syntactical theory and it only captures elements of syntax implicitly via the use of an n-gram language model in the noisy channel framework which cannot model long dependencies. However, the goal of a syntax-based machine translation techniques is to incorporate an explicit representation of syntax into the statistical systems to get the best out of the two worlds: high quality output while not requiring intensive human efforts (Ahmed and Hanneman 2005).

Koehn (2012) gives the following as the advantages of a Syntax-based Model:

- Reordering for syntactic reasons _ – e.g., move German object to end of sentence
- Better explanation for function words_ – e.g., prepositions, determiners
- Conditioning to syntactically related words_ – translation of verb may depend on subject or object
- Use of syntactic language models

Ahmed and Hanneman (2005:4) report that Koehn, Och, and Marcu (2003) are of the opinion that syntax is detrimental and does not boast the performance, but decreases the accuracy.

In the words of Koehn, Och, and Marcu (2003:7) they explain that:

> Straight-forward syntactic models that map constituents into constituents fail to account for important phrase alignments. As a consequence, straight-forward syntax-based mappings do not lead to better translations than unmotivated phrase mappings. This is a challenge for syntactic translation models. It matters how phrases are extracted. The results suggest that choosing the right alignment heuristic is more important than which model is used to create the initial word alignments.

Ahmed and Hanneman (2005) criticise the position of Koehn, Och, and Marcu (2003) of being biased. They also view that syntax was loosely defined. Ahmed and Hanneman (2005) believe that one possible characterization of syntax in the context of MT is a method used to generate constituents and model their movement across languages. Hence they explain various formal grammars that can be used for the exercise as enumerated below:

- Inversion Transduction Grammars
- Synchronous Context-free Grammars
- Multitext Grammars
- Synchronous Tree-Adjoining Grammars

- Synchronous Tree-substitution Grammars and
- Probabilistic Interpretations

Space and time will not allow us to review any of the grammars above rather we shall move to review another model of Phrase Based SMT that makes use of one of the formal grammar mentioned.

2.4.4 Hierarchical Phrase-Based SMT

One of the syntax support for Phrase-Based SMT is the hierarchical phrase based translation MT. Hassan (2009:23) report that Chiang (2005) was the first work to demonstrate any improvement when adding hierarchical structure to phrase based SMT. He explains that the approach uses hierarchical phrase transduction probabilities to handle a range of reordering phenomena in the correct fashion. He goes further to explain that Chiang (2005) proposes a generalised form of the phrase where a synchronous context-free grammar is used to provide the ability of inserting a sub-phrase into a larger phrase. The derived transduction grammar does not rely on any linguistic annotations or assumptions so that the 'syntax' induced is not linguistically motivated and does not necessarily capture grammatical preferences in the output target sentence. He stresses that all the phrases have a single generalisation category and, thus, each phrase can be substituted for any other phrase and an n-gram language model is used to judge the resulting phrases. This approach requires a chart-based decoding which has much more computational cost than the beam search decoding used for the phrase-based SMT. He points out that Chiang (2005) uses a small language model to avoid the complex search requirement when adding a large n-gram language model.

To Almagbout (2012) hierarchical phrase-based (HPB) SMT is a tree-based SMT model which extracts a synchronous Context Free Grammar (SCFG) from parallel corpus without using a syntactic annotation. He explains that in HPB SMT, the SCFG is used to parse the source sentence while generating the target translation. He further states that:

- HPB SMT framework uses the log-linear model to combine the various component models which participate in the calculation of translation probability and performs tuning through Minimum Error Rate Training just like the phrase based SMT framework.

- HPB SMT like PB SMT uses the beam search decoding algorithm combined with chart decoding in addition to rescoring techniques originally developed for PB SMT.
- The pipeline of the HPB SMT system shares many similarities with the pipeline of the PB SMT system.

Chiang (2005) proposes a set of techniques which limit the size of the extracted grammar as enumerated by Almagbout (2012) below:

- The length of initial phrase is limited to 10 words on the source side. The number of words and nonterminals in the source side is limited to 5.
- Adjacent nonterminals are prohibited in the source side of the rule in order to avoid superious ambiguity.
- The number of nonterminals on each side of the rule is limited to 2.
- Unaligned words are prohibited at the boundaries of the initial phrase.
- The rule should have at least one pair of aligned words (e.g. every rule must contain at leasts one terminal symbol).

He (Almagbout 2012:23) then concludes that in contrast to PB SMT, HPB SMT does not need a separate phrase reordering model given the availability of hierarchical phrases which capture highly lexicalised phrase reordering, whereas continuous phrase only capture word reordering. Initial rules, which are identical to the phrase pair used in PB SMT, give HPB SMT the power of continuous phrases too. He went further to say that hierarchical rules enable the translation of discontinuous phrases because it is important to capture many linguistic phenomena which are not directly (or even impossible) to be captured by PB SMT. For further details of this approach see Cai, Lu, Liun (2009), Chiang (2005 and 2007), Hayashi, Tsukada, Katsuhito, Duh, Yamamoto (2010).

2.5 Challenges of MT

The challenges of MT has been identified and classified. We consider these challenges based on the authors. Odoje and Akinola (2013) make reference to Bar-Hillel (1953) who identified the challenges of MT from the linguistic point of view. He identifies four challenges:

- Operational Syntax;

- Inter-translatability of natural languages;
- Idioms, and
- Universal syntactic categories.

Bar-Hillel explicates that one of the decisive steps in certain methods of MT is the determination of the syntactic structure of any given sentence in the *source language* to a required degree of explicitness. He views a machine as an utterly moronic student without the slightest knowledge of either the *source language* or the *target language* syntactic categories. What the machine can do is *matching* the given text or any part of it with any of a number of lists presented to it, and *counting.* Bar-Hillel is of the view that the linguist has to be provided with an *Operational Syntax,* which explains what to do first as well as what to do at the *n-th* step depending on the outcomes of the preceding *n-1* steps (preferably, of the (n-l)th step only). To Bar-Hillel, no sufficiently complete operational syntax of any language has thus far been produced, mainly because the importance of such syntax has not been recognized. Although this importance is highlighted by MT, it extends far beyond the reaches of the specific application. The preparation of an operational syntax for any or all languages is, in his opinion, a task which should prove highly rewarding even for the most theoretically minded linguist.

Bar-Hillel went further that inter-translatability of natural languages is highly ambiguous due to the ambiguity of both "inter-translatability" and "natural language" (Bar-Hillel 1953: 219). He explains natural language from two mutually exclusive senses: close and open language. A closed language is one whose rules, both of the syntactic and the semantic nature, derive from the behaviour of its users at a certain time according to principles which, at least in theory, are well understood, rigid and unalterable. This implies, also, a fixed and unexpandable vocabulary. He opines that not only in the sense that the prospective translator would be unable to complete his task in a reasonable time but that a completion of the task would be theoretically impossible. For the open language, he mentioned that translation would be an easy task if the possibility of extension is taken seriously, or rather hyper-seriously, the task would become not only easy but utterly trivial.

This result, which should have some debunking value, was obtained even without taking into account the ambiguity of the term "*inter-translatability.*" It is difficult to know in what sense this term and its cognates were understood by those who used them in connection with the problem. If they had in mind a relation that is stronger than sentence-by sentence- translatability, they were probably wrong in every interpretation except the utterly trivial one mentioned above. Under no restricted extensibility does it seem plausible that, in general, smaller units than sentences will turn out to be uniquely translatable. It is not even clear that sentences are large enough units.

Bar-Hillel (1953: 221) emphases that among the obvious difficulties that also arise when considering MT is the treatment of idioms. Somehow, one can envisage how a machine could proceed in a kind of word-by-word translation but it is exactly this type of translation which collapses when confronted with an idiom which by definition is an expression in the usage of a language, that is peculiar to itself either in grammatical construction or in having a meaning which cannot be derived as a whole from the conjoined meanings of its elements. He concludes that from the meaning of the term, “idiom”, with respect to a target language and a set of translation rules, it follows that no idiom can be satisfactorily translated into target language by a machine that follows the rules. Therefore, the only method of mechanically translating idioms is—not to have idioms at all.

The last of Bar-Hillel categorization of challenges of MT is Universal Syntactic Categories. He is of the opinion that a syntactic categorization cannot be universally applicable to all languages though he agrees that there are some categories that are common to all languages such as proper names, or at least, expressions which could be considered as proper names under some slight pressure.

It should be remarked that Bar-Hillel’s argument reduces MT to merely a rule-based approach and word translation without considering other major technical and socal challenges. It must be noted that events have taken over most of his claims. Other approaches, especially Statistical Machine Translation, do not consider lexical or syntactic categorization for translation; the approach is basically mathematical where no recall is made to any linguistic information. Using the Chomskian grammar, the question of Operational Syntax and Universal Syntactic Categories is now solved. As much as the universality of language cannot be denied, the concern should be how to capture the universal properties of natural language and the unique feature of the individual languages for the purpose of machine translation.

On inter-translatability, the problem is not limited to the machine alone but the problem of language and translation in general. A language is a reflection of the culture it represents and no two cultures are exactly the same; therefore, translating languages of two different cultures may not be very easy. For instance Osundare (1995) observes that it is problematic to represent the culture of a language in another language. So, if translation could be that difficult for humans, how much more a machine?

Och (2006) identifies four major challenges of MT different from Bar-Hillel's but his focus is on Statistical Machine Translation (SMT). He notes that collecting and using huge amounts of data for achieving optimal MT quality is a problem. Train models like language models, translation models etc. are done on hundreds of millions of words that require very large computational resources, resulting in higher overhead on computation time, the number of steps necessary to solve a problem, memory space and the amount of storage needed to solve the problem. He lists machine learning problems as another challenge. This has to do with structured prediction on very large amounts of training data, and in particular, the use of discriminative training techniques based on millions of features seems to be promising but would also require even greater computational resources.

The next challenge he notices is on evaluation of MT translation. He explains that the performance improvements achieved by MT systems based on very large amounts of data have been very significant – so significant that existing automatic evaluation metrics (e.g. BLEU) have a hard time distinguishing MT output and human translation output on some standard data sets. Hence, the need to find new MT evaluation metrics that can drive progress in the coming years. These new metrics like BLEU must be fully automated once initial human translations/judgments have been collected.

The last challenge he points out has to do with the interdisciplinary approach to MT. He explains that another debated question in MT is its relationship to Computational Linguistics (CL) and Natural Language Processing (NLP) research. Currently, the best data-driven MT systems do not employ NLP tools such as linguistic parsers, parts of speech taggers or explicit word sense disambiguation, and there have been very few success stories in integrating those components.

Lopez (2008) too identifies four problems in building a functional SMT system. Like Och (2006), Lopez's focus is on SMT and no reference is made to any African language. None of the aforementioned scholars considered the peculiarities of African languages during their classification. However, the challenges of MT in relation to African languages have been identified but they are considered in isolation with the existing challenges. For example, Adegbola (2008) and Odoje (2011) reiterate challenges faced while developing MT for African language but it was based on the experience of Linguistic Rule Base MT. Also, Odoje (2012) identifies some of the challenges of MT beyond the rule-base approach but they were not categories. Hence, the next section focuses on challenges of MT as it relates to African Languages.

2.6 Challenges of MT as it Relates with African Languages

Digital resources for Yorùbá content are scarce; as a result many MT in the language opt for the rule-based MT system (Awofolu 2002, Odoje 2010, Odoje 2012, Odoje and Akinola 2013). Odoje (2010) enumerates the challenges of building a rule-based MT for African languages, such as the cumbersome rules needed to capture very simple sentence as well as translation; the inability to capture figurative expressions and idioms, the failure to incorporate tone manipulation among others.

Odoje and Akinola (2013) group the challenges of developing a Yorùbá-English SMT into two: the social-cultural and the technical challenges. They explain that social and cultural features of the languages involved should be fully integrated such as word order, orthography, diacritic symbols, perception and behaviour of people towards their language. Technical challeenges are view inline with keyboard support and diacrirics; funding; operating system; MT tools and availability of corpus.

2.7 Machine Translation Evaluation

A hotly debated topic in machine translation is evaluation (Koehn 2010:24). This seems so because there are many valid translations for each input sentence. Therefore, there is a need to assess the quality of machine translation systems qualitatively, or at least a way to be able to tell if one system is better than another or if a change in a system leads to an improvement (see Koehn 2004, 2007, 2010). There are two broad means of evaluating MT systems: a way is human judgment (also known as manual evaluation) which has been adjudged to be extensive but expensive, time consuming, involving human labour which cannot be used again (unreuseable) (Papineni, et al 2001; Papineni, et al 2002; Koehn 2004, 2007, 2010; Callison-Burch et al 2009; Padó et al 2010). Human judges are to assess the adequacy (preservation of meaning) and fluency of machine translation output, or to rank different translations of an individual sentence.

The other way is the use of automatic evaluation metrics. Koehn (2010:25) explains that typically, automatic evaluation metrics compare the output of machine translation systems against human translations. While common metrics measure the overlap in words and word sequences, as well as word order; advanced metrics also take synonyms, morphological variation, or preservation of syntactic relations into account. They are however evaluated by their correlation to human judgment. There are several automatic evaluation metrics but prominent among them are: Bilingual Evaluation Understudy (BLEU), and METEOR among others.

2.7.1 Goals of Automatic Evaluation Metrics

Human evaluation no doubt is expensive in terms of cost and time so much so that it is also believed to be inconsistent and the energy is un-reusable at the other time. Papineni, Roukos, Ward and Zhu (2001) opine that developers would wish to benefit from an inexpensive automatic evaluation that is quick, language-independent, and correlates highly with human evaluation. One of such automatic evaluation is Bilingual Evaluation Understudy (BLEU). Koehn (2010:234) enumerates the five goals of any automatic metric: being low-cost, tunability, meaningfulness, consistency and correctness. Low-cost means that the metric should be able to quickly and cheaply carry out evaluations of a new system, and a new domain since cost is the major disadvantage of evaluation metrics that include human evaluators, especially bilingual evaluators. Cost may be measured in terms of time or money spent on the evaluation. Being tunable means that the fully auotmatic metric should be used directly in the automatic system optimisation. When a metric ranks a system against another, the idea is to have a meaningful metric ensuring that any score given gives an evidence of leniency of the evaluator.

Consistency should be maintained across many dimensions. In other words, different evaluators, using the same metric should come to the same conclusions. Koehn called this the **inter-annotator agreement**. Meaningfulness refers to the fact that the evaluation on one part of the test corpus should be consistent with the evaluation on another part. If there is high fluctuation, i.e., the metric is not **stable**, this means that we would need large test corpora to ensure that the results are reliable. Lastly, the goal of any metric is to come up with correct results. In other words, to what extent do the outcomes correlate with the fluency and adequacy judgments? Koehn (2010:235) also points out other evaluation criteria which are very important in the discourse of SMT like speed, size, integration as well as domain adaptation and customization.

2.7.2 Automatic Evaluation Metrics

All automatic evaluation metrics use the same trick. Each system translation is compared against one or more human translations of the same sentence. The human translations are called reference translations (Koehn 2010:236). An automatic evaluation metric matches the output of an MT with reference translation whereby an output closer to reference translation is preferred by giving it a high score. This could be used to compare two or more MTs or evaluate improvement in an MT based on more training or alterations made on it (Lin and Och 2004).

Precision and Recall are part of important metrics used for automatic evaluation. Consider example (12) below:

12.

System A: He saw the Divine

Reference: He too saw Olu

System B: He saw Olu

You will observe that system A translates *Olu* as Divine which is right in some context but violate translation rule that personal names should not be translated and because *Divine* is completely different from what is obtainable in the reference translation, *Divine* is not considered as the translation of Olu. So, only two out of the total words of four are shared hence the ration of system A compared to reference translation is 50%. This is termed as precision. System B has 100% precision in that all its words are shared in the reference translation except that *too* is missing in the system B's output which is considered to be very important. However, how many of the words a system generate are correct? We need a metric called recall. This is to divide the number of correct words with the length of reference translation, instead of the length of the system output:

$$\text{Precision} = \frac{correct}{output-length}$$

$$\text{Recall} = \frac{correct}{reference-length}$$

Koehn (2010:237) reveals that both precision and recall can be deliberately tricked. We can have a situation where precision is very high and recall is low or recall is high and precision is very low. In as much as we do not want to output wrong words, at the same time we do not want to miss anything either. A common way to combine precision and recall is using the **f-measure**, which is defined as the harmonic mean of the two metrics:

$$\text{f-measure} = \frac{precision \times recall}{(precision + recall)/2}$$

but Koehn (2010) used $\text{f-measure} = \dfrac{correct}{(output - length + reference - length)/2}$

Other automatic evaluation metrics Koehn (2010) explains are Position-Independent Error Rate (PER) and Word Error Rate (WER). Position-Independent Error Rate is similar to recall in that it uses reference translation as a divisor. Because it is an error rate, mismatches are measured not matched. The metric considers superfluous words that are needed to be deleted to overcome the problem of long translations:

$$\text{PER} = 1\text{-} \frac{correct - \max(0, output - length - reference - length)}{reference - length}$$

On the other hand, Word Error Rate (WER) is borrowed from speech recognition and it take into account word order. WER employs the Levenshtein distance, which is defined as the minimum number of editing steps – insertions, deletions, and substitutions – needed to match two sequences. The task of finding the minimum number of editing steps can be seen as finding the optimal path through the word alignment matrix of output sentence (across) and reference translation (down); using a dynamic programming approach. WER normalizes the number of editing steps by the length of the reference translation:

$$\text{WER} = \frac{substitution + insertion + deletions}{reference - length}$$

Koehn (2010:239) points out that there could be a perfect translation but with different word order to the reference translation which may be marked with a very high word error rate going by the WER. This, he said is harsh if we intend to meet up with the requirement of matching word in order.

2.7.3 Bilingual Evaluation Understudy (BLEU)

BLEU is one of the current and popular automatic evaluations which have elegant solution to the role of word order. It works similarly to position-independent word error rate but considers matches of larger n-grams with the reference translation. When the n-gram matches, the n-gram precision can be computed, that is the ratio of correct n-grams of a certain order *n* in relation to the total number of generated n-grams of that order (see Panineni, Roukos, Ward and Zhu 2002; Koehn 2010; Wolk and Marasek 2014). Koehn (2010:240) defines BLEU metric as:

$$\text{BLEU-n} = \text{brevity} - \text{penalty exp} \sum_{i-1}^{n} \lambda_i \log \text{precision}_i$$

$$\text{brevity} - \text{penalty} = \min\left(1, \frac{output-length}{reference-length}\right)$$

He stresses that the problem with precision-based metrics is that no penalty for dropping words which is addressed by BLUE with a brevity penalty. The penalty therefore reduces the score if the output is too short. However, the maximum order *n* for n-grams to be matched is typically set to 4. The metric is then called BLEU-4. So, the weights λ_i for the different precisions are typically set to 1 which simplifies the BLEU-4 formula to

$$\text{BLEU-4} = \min\left(1, \frac{output-length}{reference-length}\right)\prod_{i-1}^{4} precision_i$$

He points out that BLEU score is 0 if any of the n-gram precision is 0, meaning that no n-grams of any particular length are matched anywhere in the output. Since n-gram precision of 0 especially for 4-gram often occur on the sentence level, BLEU scores are commonly computed over the entire test set. He equally made mention of multiple reference translation which is another innovation of the BLEU score. He maintains that if multiple human reference translations are used, it is more likely that all acceptable translations of ambiguous parts of the sentences will show up. It should be noted that multiple reference translations complicate the issue of reference length. So, the closest length of each output sentence is determined and taken as the reference length. If two reference lengths are equally close, but one is shorter and the other is longer, the shorter one is taken. For instance, given an output length of 10 and lengths of reference sentences 8, 9, 11, and 15, the reference length for that sentence is 9 (both 9 and 11 are equally close, but 9 is smaller). However, Papineni, Roukos, Ward and Zhu (2002:5) had earlier warned that it is important to note that the more reference translation per sentence, the higher the score. Thus, one must be cautious making even “rough” comparisons on evaluations with different numbers of reference translation.

2.7.4 Metric for Evaluation of Translation with Explicit Ordering (METEOR)

METEOR was designed to explicitly address several observed weaknesses in IBM's BLEU metric. Banerjee and Lavie (2005) are of the opinion that BLEU uses n-gram precision which is calculated separately for each n-gram order and are combined via geometric averaging. This means that BLEU does not take recall into account directly. To them, recall is extremely important for assessing the quality of MT output as it reflects to what degree the translation covers the entire content of the translated sentence. They maintain that BLEU does not use recall because the notion of recall is unclear when matching simultaneously against a set of reference translation. They believe that the brevity penalty in BLEU does not adequately compensate for lack of recall.

They suggest that an explicit measure for level of grammaticality (or word order) can better account for the importance of grammaticality as a factor in MT metric and the result would be better in correlation with human judgment. Hence, n-gram counts in BLEU do not require an explicit word-to-word matching which can result in counting incorrect "matches", particularly common functors (function words).

In the same vein though from another perspective, Koehn (2010:228) also flaws BLEU that it gives no credit to near matches in term of stemming, synonyms or semantically closely related words and multiple reference may involve choice of word which BLEU may not be able to score. He points out that METEOR on the other hand incorporates the use of stemming and synonyms by first matching stems and finally semantic classes. The latter are determined using Wordnet, a popular list of English words that also have near equivalences in other languages. He also mentions the main drawback of METEOR as its method and formula for computing a score being much more complicated than BLEU's. The matching process involves computationally expensive word alignment. There are many more parameters such as the relation weight of recall to precision, and the weight for stemming or synonym matches that have to be tuned.

Lavie (2010) also explains that on the average, hypotheses are scored at a rate of 500 segments per second per CPU core and METEOR consistently demonstrates a high correlation with human judgments in independent evaluations such as EMNLP WMT 2011 and NIST Metrics MATR 2010.

2.7.5 NIST Metric

Having considered BLEU and METEOR, other automatic evaluation metrics that will be considered are NIST and TER. As stated on the NIST website, MT evaluation series started in 2001 as part of the DARPA TIDES (Translingual Information Detection, Extraction) programme. Beginning with the 2006 evaluation, the evaluations have been driven and coordinated by NIST (National Institute of Standards Technology) as NIST OpenMT. These evaluations provide an important contribution to the direction of research efforts and the calibration of technical capabilities in MT. The Open MT evaluations are intended to be of interest to all researchers working on the general problem of automatic translation between human languages. To this end, they are designed to be simple, to focus on core technology issues and to be fully supported (http://www.nist.gov/itl/iad/mig/openmt.cfm). There has been progress test from the OpenMT 2008, 2009, 2012 evaluations with new source data created by humans based on the English reference translation. The NIST Open Machine Translation 2015 Evaluation (OpenMT15) will be implemented according to the OpenMT15 evaluation plan. It took place in February 2015, followed by a workshop in May 2015. The highlights of the workshop include:

- evaluation on informal data genres (SMS/chat, telephone conversations) for Arabic-to-English and Chinese-to-English;
- inclusion of audio input track; and
- explanation of common MT measurement techniques on these informal data genres

And in August 2015, there was an official release of the results of evaluations.

2.7.6 Translation Edit (Error) Rate

Snover, Dorr, Schwartz, Micciulla, and Makhoul (2006) report that GALE (Olive 2005) (Global Autonomous Language Exploitation) research programme introduced a new error measure called Translation Edit Rate (TER). TER was originally designed to count the number of edits (including phrasal shift) performed by a human to change a hypothesis so that it is both fluent and has the correct meaning. TER is however defined by Snover et al (2006) as the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references normalised by the average length of the references. Since the concern is the minimum number of edits needed to modify the hypothesis, only the number of edits to the closest reference is measured (as measured by TER score).

$$\text{TER} = \frac{\# of edit}{average \# of reference words}$$

Possible edits include the insertion, deletion and substitution of single words as well as shifts of word sequences. All edits, including shifts of any number of word by any distance, have equal cost. And punctuation token are treated as normal words and mis-capitalisation is counted as an edit.

Snover et al (2006) conclude that HTER is expensive in that it requires approximately 3 to 7 minutes per sentence for a human to annotate. They recommend that fewer references would likely be adequate, reducing the cost of the method, relative to methods that require many reference translations. They are of the view that HTER is not suitable for use in the development cycle of an MT system. Although it could be employed on periodic basis, it appears to be a possible substitute for subjective human judgments of MT quality.

They equally maintain that TER is easy to explain to people outside the MT community (i.e., the amount of work needed to correct the translations). Both TER and HTER appear to be good predictors of human judgments of translation quality. In addition, HTER may represent a method of capturing human judgments about translation quality without the need for noisy subjective judgments. The automatic TER score with 4 references correlates as well with a single human judgment as another human judgment does while the scores with a human in the loop such as HTER, correlate significantly better with a human judgment than a second human judgment does. This confirms that if humans are to be used to judge the quality of MT output, this should be done by creating a new reference and counting errors, rather than by making subjective judgments.

2.7.7 Linguistics and Automatic Evaluation of MT

Lavie et.al (2007) report that automatic metrics for MT evaluation have been receiving increasing attention over the past five years. These scholars reiterate that such metrics are critical tools for current and future MT research as they allow research teams to guide the development of their system based on frequent concrete performance evaluations. They emphasize that the models used by MT systems today and probably in the future contain a variety of parameters that need to be tuned for optimal performance. They opine that as translation quality improves, attention should be given to small but sensitive differences at the sentence level to further achieve better translation qualities. The following excerpt from Lavie (2007:10) puts this in perspective:

> As MT systems improve and achieve high level of translation quality, it becomes ever more important to have evaluation metrics that are sensitive to small differences between translations at the sentence-level, so that minor improvements can still be detected, concrete translation errors can be isolated and identified and system parameters can be optimized to truly achieve the best translation performance...

This implies that MT and its evaluation metrics still do not take the so-called minor but complex and complicated words into account while translating or evaluating. This is the possition of MT critics. For example, Bar-Hellie (1954) maintains that MT, especially Statistical Machine Translation (SMT)[6], limits human natural language to counting and merging. Each word's features/properties are checked before merging takes place in human natural language. SMT on the contrary, does not reckon with the inherent properties of words in the lexicon before they enter computation. This makes it easy for critics to assume that no matter the success recorded by SMT, there would still be some intricacies such as nuances, contextual meaning, semantic extension, etc, yet to be covered.

The blame is not entirely that of SMT practitioners; linguists and translators also have a huge part in the blame in that they left the discourse entirely to SMT practitioners (Way 2012:9). Way (2012) provides two suggestions to reduce the impending challenge in the discourse of SMT as shown in the extract below:

> ... as SMT became the principal way of doing MT, this conciliatory tone soon changed, to the point today where many people who want to understand have been left so far behind that they feel that it is impossible to ever catch up. We expressed that the view that linguists and translators have to share the blame in allowing the field to move almost entirely in the statistical direction, especially when the seminal IBM papers very much left the door open for collaboration with the

[6] Statistical machine translation (SMT) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora.

> linguistic community. However, in our view SMT research will soon have to alter their position, if the use of syntax (and later, once a further ceiling has been reached, semantics) is to become mainstream in today's model. These syntactic improvement have largely come about from those practitioners with wider background than is the norm in SMT. Those without a linguistic background, then, appear to have two choices: (i) to attempt to include the linguists, so that they may be of help; (ii) to continue to exclude linguists, while at the same time trying to make sense out of their writings...

To corroborate Way's point of view, one wonders the volume of parallel corpora needed for a MT to translate the Yorùbá verb, *pa,* with its different meanings as contexts of usage demand. The examples (13) below show how *pa* can have different senses/meanings, depending on contexts and usages which the available corpora may not accommodate.

13 a Adé pa ejò
Adé kill snake
‘Ade killed a snake’

b Adé pa irọ́
Ade tell lie
‘Ade lied’

c Adé pa ilẹ̀
Ade clear bush
‘Ade cleared the bush’

d Adé pa ilé ní aró
Adé paint house prep colour
‘Ade painted the house’

e Adé pa èkùrọ́
Ade crack palmnut
‘Ade cracked the palmnut’

f Adé pa ojú dé[7]
Ade v eyes v
‘Ade closed his eyes’

g Ojò pa Adé
Rain beat Ade
‘Rain beat Ade’

[7] padé is a split verb meaning to close

h Adé pa itu ọdẹ
Ade perform wonder hunter
‘Ade perfomed wonders’

i Adé pa owó rẹpẹtẹ
Ade make money plenty
‘Ade made lots of money’

j Adé fi ìja pa ẹẹ́ta pẹ̀lú Ṣadé[8]
Ade use fight perform three prep Sade
‘Ade fought Sade’

k Adé fi òróró pa orí
Ade use oil rub head
‘Ade creamed/rubbed/ anointed himself with oil’

l Adé pa àsẹ fún wa
Ade issue command prep us
‘Ade commanded us’

m Adé pa òṣé sí ọ̀rọ̀ rẹ̀
Ade made hiss prep word pro
‘Ade hissed at his word’

n Adé pa kuuru sí wọn
Ade perform rush prep they
‘Ade rushed at them’

o Adé a máa pa ariwo
Ade aspect make noise
Ade makes noise

p Ọ̀rọ̀ pa èsì jẹ[9]
Word kill response eat
‘No comment’

r Adé pa àtẹ́wọ́
Adé clap palm
‘Ade clapped’

s Adé pa ilẹ̀ mọ́
Ade clear ground clean
‘Adé prepared/ cleaned up the surroundings’

[8] pa in example 1j is used in its idomatic and figurative sense as such its meaning is different from that of everyday usage.
[9] Example 1p is an idiomatic expression.

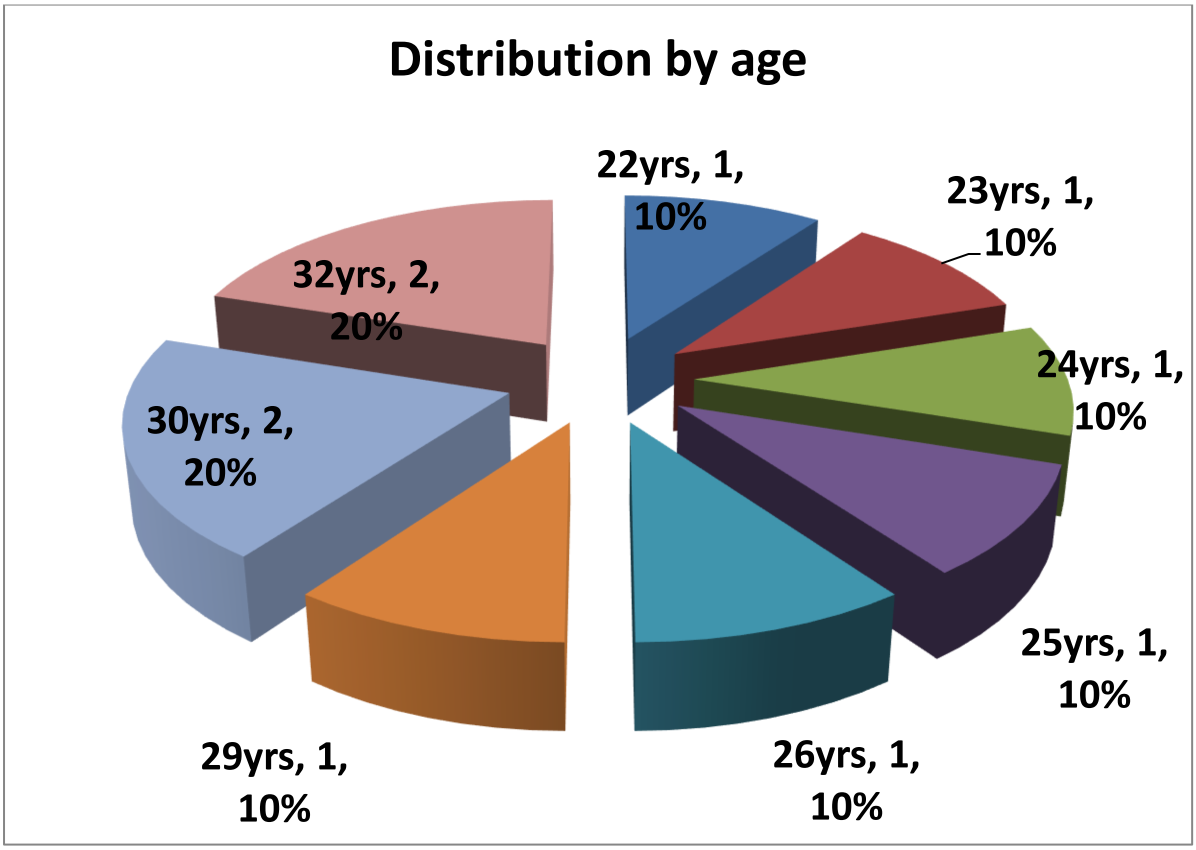
Examples like those in (13) above show how difficult it is to achieve translation via machine, whether rule or statistical based. While rule-based MT limits translation to the structures of a language and is faced with structural ambiguity issues among other challenges, the statistical-based machine translation is devoid of usage in context because no corpus of any language is big enough to have all the possible words of the language and its contextual usage in a print whether literary or other forms. For example *pa* is used 304 times[10] in *Igbó Olódùmarè* which has 1744 sentences. Fágúnwà's *Igbó Olódùmarè* uses *pa* in all contexts above except those illustrated by example (13 b,c,e,and i). This means that the SMT that uses the novel as its training corpus will not be able to translate the said examples in the context of usage as mentioned above. This is even better compared to Folajimi and Omonayi (2012) who use Genesis (the first book of the Bible) as their training corpus whereby, the whole of Yorùbá Bible has 7 out of 17 contextual usages exemplified in (13) above (i.e. a, b, f, g, k, l, and o). As much as we agree with Jurafsky and Martin (2009) that it may be difficult to acquire the legal right to fiction as well as translate literal sentences which may require compromise, yet we are of the opinion that these literary texts could serve as bases for resource-scarce languages like Yoruba to start the building of SMT pending when there would be enough resources. We also believe that neutral/non-religious literary materials will yield better results than the Bible that is restricted to religious vocabulary, although, this suggestion is just one of the numerous ways of overcoming language modelling in MT.

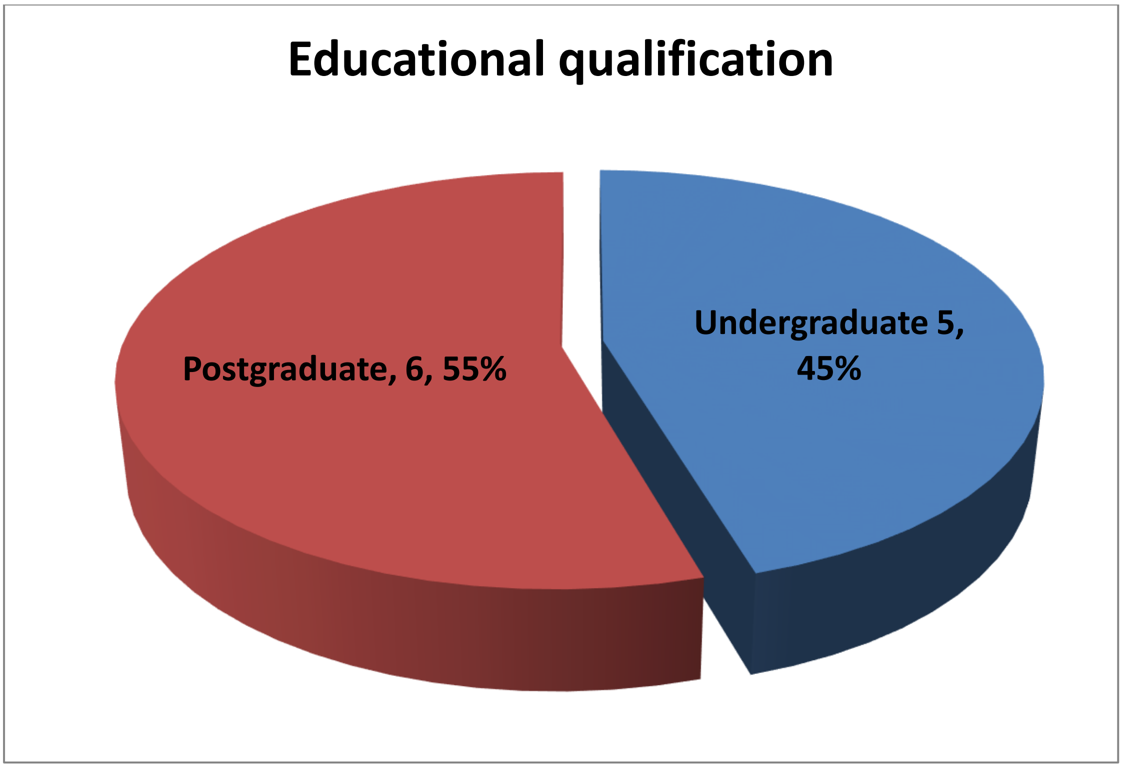
2.8 Human Evaluation of Google Translate

As of today, Google is the only available bidirectional free MT application for Yoruba-English translation task. This study subjects its translations to human evaluation because, human evaluation is adjudged to be the best and most accurate even though it too has its challenges. The Ibadan and Akungba Structured Sentence Paradigm (SSP) were employed to test the extent at which it translates the language pair and its bidirectional claim of translation. While Ibadan SSP is drafted in English, Akungba SSP is in Yoruba. The SSPs were translated by human and Google Translate and their

[10] Going by Awobuluyi (2013:88) explanation of the verb form in Yorùbá, we consider words like *pàdé, padà, parí* and *papọ̀* as compound verbs derived from *pa* and other verbs in the language which make up 304 occurences of the verb.

translations were then subjected to human evaluation. Twenty (20) translation practitioners in the Department of Linguistics and African Languages, University of Ibadan were contacted to evaluate the translations but only eleven (11) responded. The charts below reflect the age and educational qualifications of the respondents





According to Koehn (2010:232), one of the common approaches in human evaluation is the use of graded scale. The graded scales are based on fluency and adequacy. Fluency indicates that the output is a fluent target language involving both grammatical correctness and idiomatic word choices. Adequacy is concerned with the output conveying the same meaning as the input sentence without losing, adding or distorting any part of the message. The graded scale is given in Table (1) below:

Table 1

Adequacy	
5	All meaning
4	Most meaning
3	Much meaning
2	Little meaning
1	None

Fluency	
5	Flawless English
4	Good English
3	None-native English
2	Disfluent English
1	Incomprehensible

Table 1 informs the scale which the evaluators used in evaluating both machine and human translations. Table 2 shows each evaluator's rating of both human and machine translations and their average scores (the evaluator's rating divided by the number of items rated e.g 247/160 = 1.544).

Table 2

Score of fluency of interpretation for Google Translate	**Average Score for fluency of interpretation for Google Translate**	**Score of adequacy of translation for Google Translate**	**Average Score of adequacy of translation for Google Translate**	**Score of fluency of interpretation for human translation**	**Average Score for fluency of interpretation for human translation**	**Score of adequacy of translation for human translation**	**Average Score of adequacy y of translation for human translation**
247	1.544	228	1.425	735	4.594	705	4.406
323	2.019	257	1.606	776	4.85	756	4.725
218	1.363	216	1.35	731	4.569	724	4.525
201	1.256	198	1.238	620	3.875	612	3.825
245	1.531	199	1.244	692	4.325	685	4.281
219	1.369	217	1.356	629	3.931	629	3.931
213	1.331	209	1.306	620	3.875	622	3.888
234	1.463	230	1.438	575	3.594	579	3.619
195	1.219	181	1.131	633	3.956	635	3.869
264	1.65	270	1.688	705	4.406	709	4.431
190	1.188	192	1.2	624	3.9	623	3.894
2549	**15.933**	**2205**	**14.982**	**7340**	**45.875**	**7279**	**45.394**
	Mean Score 15.933/11= 1.449		**Mean Score 14.982/11= 1.362**		**Mean Score 45.875/11= 4.171**		**Mean Score 45.394/11= 4.127**

Mean is an arithmetic average of scores, calculated by adding all the scores, divided by total number of scores. This helps to decide which group has higher performance in cases where means are compared. Table (2) shows that the mean score of the accuracy and fluency of both Google Translate as well as human translation. The mean score of Google Translate adequacy is 1.362 which is approximately equal to 1. This implies that using the ranking in Table (1), the adequacy of Google Translate falls into category 1 which is very poor. This then means that the adequacy of Google Translate translating Yoruba to English is very poor. The mean score of human adequacy is 4.127 which approximately equals 4. From the rank of our scoring, 4 falls into the category of very good, meaning that human translation is more adequate than Google Translate.

The result of the fluency shows that the mean score of MT is 1.449 which approximately equal to 2. From the rank of our scoring about the fluency of translation in Table (1) above, 2 falls into the category of poor, implying that the fluency of computer translating Yoruba to English is poor. However, the mean score of the human translation is 4.171, which approximately equals 4. Since 4 above falls into the category of very good in the ranking of our scoring in table (1), it means that the fluency of human translation is better in comparison for possible translation to that of MT.

We also carried out a paired t-test[11] at 0.05 confidence interval to check if there is a significant difference in the means of MT and human translation. The result of the analysis is significant with p-value at 0.000. The result of the paired t-test is significant; in other words, there is a significant difference between human translation and Google Translate translations. Also, the mean of human translation (1329.00) is higher than that of Google Translate (449.73) showing that human translation is more efficient and effective than Google Translate as the tables below show:

Table 3

T-Test

[11] A T-test is a statistical examination of two population means. It examines whether two samples are different and commonly used when the variances of two normal distributions are unknown and when an experiment uses a small sample size (www.investopedia.com)

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Human Translation	1329.00	11	119.174	35.932
	Machine Translation	449.73	11	62.629	18.883

Table 4

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Human Translation & Machine Translation	11	.674	.023

Table 5

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Human Translation - Machine Translation	879.273	89.790	27.073	818.951	939.594	32.478	10	.000

We should not just conclude that human translation is much better in terms of adequacy and fluency of translation when compared with translation done by Google Translate. At this point, the concern should be the explanation of the occurred errors.. For example, Google Translate translates *olú* as *emperor, divine* and *capital.* We observe that the error is not entirely the machine's rather the training corpus is responsible. We are not sure of the Google Translate source of training corpus so we assume that because the general translation of *capital city* is *olú-ìlú, Olú* was translated as *capital.* Also, *divine/divinity* could be interpreted as *Olú ọ̀run/Òrìsà/Olúwa* which may inform the machine to tanslate *Olú* as d*ivine* sometimes. Also, o*lú* could mean *emperor*, depending on the context, but the general usage of Olú being a personal name does not need any translation. Therefore there could be a high frequency of olú's being translated as *emperor*, *divine* and *capital* based on the information provided in the training corpus as shown in example (14) below:

14	**Source sentence**	**Translation by Google Translate**	**Translation**
a.	Olú rí mi	My capital	Olú saw me
b.	Olú rí ẹ/ọ	Capital	Olú saw you
c.	Olú rí i	Capital found	Olú saw him/her
d.	Olú rí wọn	Their capital	Olú saw them
e.	A rí Olú	A capital	We saw Olú
f.	Ẹ rí Olú	Their capital	You(pl) saw Olú
g.	Ìwọ rí Olú	You see the Divine	You saw Olú
h.	Àwa pàápàá rí Olú	We even saw the Emperor	We too/even saw Olú
i.	Ẹ̀yin rí Olú	Emperor eggs	You(pl) saw Olú
j.	Ẹ̀yin gan-an rí Olú	Emperor eggs	You(pl) too saw Olú

The frequency of translating *olú* as *capital*, *divine* and *emperor* is not consistent. Its inconsistency needs further investigation. From the data above, whenever olu is used with short pronouns especially in the object position, *olu* is translated as *capital.* When used with pronominal, then, the choice is between *emperor* and *divine.*

14 a,b,c,d,e,f,i,and j show the translations of the Google Translate substituting sentences for phrases. It is observed that the translations do not convey, in any way, the meaning of the source language. Only 14 g and h show a reasonable level of translation except that o*lú* is mis-translate in them.

The translations also show that the Google Translate failed to learn Yorùbá with its tones as well as its manipulations (see Owolabi 2013). (14 i and j) show that the Google Translate did not notice the difference between egg (ẹyin) and the pronominal (ẹ̀yin). Hence it translates the plural third person pronominal for egg.

However, we need to point out that some complex sentences like (15) below were translated appropriately.

15	**Source Sentence**	**Translation by Machine**
a.	Kò yé mi bí mo ṣe ṣe é mọ́	I did not understand how I did it
b.	Bóyá ó bọ́ ní àpò mi	Perhaps it was lost in my bag

Even though the emphasis in the source sentence is lost in the translation by Google Translate in (15a), it is still meaningful to a large extent. (15b) is a bit ambiguous but its translation could be one of the meanings or a necessary compromise as opined by Jurafsky and Martin (2009:875). They opine that “true translation, which is both faithful to the source language and natural as an utterance in the target language is sometimes impossible, and if you are going to go ahead and produce a translation anyway, you have to compromise”. It should be noted that the aim of MT has shifted from good, quality, direct and unedited translation to the production of the first draft meant to be edited by humans.

2.9 Human Evaluation of Google Translate of English and German

To further foreground the difficulty of translating African languages by Google Translate, we carried out another experiment like the above using English and German. The result proves that, aside the fact that African languages are resource scarce languages, having a few materials necessary for machine training for statistical model like Google Translate, difference in sentence is another major challenge in the translation.

It is noteworthy to state that German and English belong to the same family. They both belong to the Germanic branch of the Indo-European language family; hence, they share many features. Apart from verb conjugation, modal verb forcing the other verbs to be at the sentence-final position and their infinitive forms which are significant areas of difference bewtten the languages, it could be concluded that what is left in the translation process is word substitution.

A Nigerian studying German in a German language school was engaged to translate SSP to German. This Nigerian is in Germany as at the time of this research and he had devoted four month to the language study. He was rated A2[12] as at the time the translation is done. We equally translated SSP to German through Google Translate. Eleven (11) human evaluators who are bilingual Germans participated in the human evaluation process. Below are information on the age, educational qualification and marital status of the participants:

Table 6

Distribution by age	**Frequency**	**Percentage (%)**
Less than 25years	6	54.54
25 years and above	5	45.45
Total	**11**	**100 %**

[12] A2 is the second level of six levels used broadly by the Europeans to ascertain the proficiency of a language learner. This level is synonymous to an elementary level with the following features: Can understand sentences and frequently used expressions related to areas of most immediate relevance. Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of their background, immediate environment and matters in areas of immediate need.

Table 7

Marital status	**Frequency**	**Percentage (%)**
Single	9	81.81
Married	2	18.18
Total	**11**	**100 %**

Table 8

Educational Level of Respondents	**Frequency**	**Percentage (%)**
Undergraduate	8	72.74
Postgraduate	2	18.2
Others	1	9.1
Total	**11**	**100%**

Table (9) shows the means of their evaluation

Table 9

S/N	Score of fluency of interpretation for Google translation	Average Score for fluency of interpretation for Google translation	Score of Adequacy of translation for Google translation	Average Score of Adequacy of translation for Google translation	Score of fluency of interpretation for human translation	Average Score for fluency of interpretation for human translation	Score of Adequacy of translation for human translation	Average Score of Adequacy of translation for human translation
1.	326	2.884956	417	3.690265	486	4.300885	502	4.442478
2.	247	2.185841	281	2.486726	424	3.752212	480	4.247787
3.	344	3.044248	344	3.044248	505	4.469027	505	4.469027
4.	265	2.345133	325	2.876106	427	3.778761	460	4.070796
5.	290	2.566372	296	2.619469	465	4.115044	463	4.097345
6.	433	3.831858	436	3.858407	527	4.663717	527	4.663717
7.	366	3.238938	321	2.840708	441	3.902655	419	3.707965
8.	301	2.663717	294	2.60177	481	4.256637	493	4.362832
9.	283	2.504425	327	2.893805	456	4.035398	462	4.088496
10.	359	3.176991	375	3.318584	501	4.433628	512	4.530973
11.	402	3.557522	329	2.911505	398	3.522124	324	2.867257
	3616	**32.000001**	**3745**	**33.141593**	**5111**	**45.230088**	**5147**	**45.548673**
		MS =32.000001/11= 2.909091		MS=33.141593/11= 3.012872091		MS=45.230088/11= 4.111826182		MS=45.548673/11 = 4.140788455

From the result of the analysis above in Table (9), the mean score of the adequacy of translation by Google Translate from English to German is 3.012872091 which is approximately equal to three (3). As the ranking in Table (4) about scoring the adequacy of translation shows, three falls into the category of good. This implies that the adequacy of Google Translate is good. The mean score of human translation adequacy is 4.140788455, which is approximately equal to 4. 4 in the Table (4) ranking falls into the category of very good, this implies that human translation is more adequate in translation than Google Translate.

The mean score of the fluency in the translation of Google Translate is 2.909091 which is approximately equal to three. Three from Table (4) shows that the translation fluency of Google Translate is good while that of human translation is very good with 4.111826182 as the mean score.

Paired T- test was also carried out at 0.05 confidence interval to check if there is a significant difference between the means of Google Translate and human translation. The result of the analysis is significant with p-value 0.000, the result of the paired t-test is significant for showing the significant difference between human translation and Google Translate. Also the mean of human translation accuracy (467.9091) is higher than that of the Google translation (340.4545), showing that human translation adequacy is more efficient and effective than the Google Translate. The mean of the human translation fluency (464.6364) is higher than that of Google Translate (328.7374), indicating that the human translation fluency is more efficient and effective than the Google Translate as shown in the tables below:

Table 10: T-test
Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Human Translation Adequacy	467.9091	11	56.54	17.04884
	Google Translate Adequacy	340.4545	11	49.79230	12.01294

Table 11

		Mean	N	Std. Deviation	Std. Error Mean
Pair 2	Human Translation Fluency	464.6364	11	39.63148	11.94934
	Google Translate Fluency	328.7273	11	58.21356	17.55205

Table 12

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Human Translation & Google Translate (Adequacy)	11	.392	.233

		N	Correlation	Sig.
Pair 2	Human Translation & Google Translate (fluency)	11	.349	.293

Table 13

Paired Samples Test

		Paired Differences					T	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Human Translation - Google Translate (Adequacy)	127.45455	58.88525	17.75457	87.89490	167.01420	7.179	10	.000
Pair 2	Human Translation -Google Translate (fluency)	135.9090	57.8834	17.45251	97.02248	174.79571	7.787	10	0.000

Even though there is a significant difference in the translation of Google Translate and human in both Yoruba and German translation of English, it is observed that the Google Translate's outputs for German experiment fluency mean score (328.7273) is higher when compared the fluency mean score (1.449) for Yoruba experiment. The same goes for adequacy; while Google Translate accuracy mean score is 3.012872091 for German experiment that for Yoruba is 1.362. With this, it is plausible to conclude that adequacy and fluency of German translation is better compared to Yoruba. Consider examples (14, 15 and 16) below:

14.

	English	**German**	**Yoruba**
A	Olu sees me	Olu sieht mich	Olú rí mi
B	We saw Olu	Wir sahen Olu	A rí Olú
C	I command Olu to go immediately	Ich befehle Olu sofot gehen	Mo pa á láṣẹ fún Olú láti lọ lẹ́sẹ̀kẹ́sẹ̀
D	Let us work so that we would have money	Lessen Sie uns zusammenarbeiten damit wir Geld haben	Jẹ kí a ṣiṣẹ́ ki a yoo ni owo

15.

Yoruba	**English**
a. Olú rí wa	Mushrooms are found
b. Adé rí wa	Crown Find us
c. Òjò pa Adé	Day off Ade
d. Ìlú wa rẹwà	our city beautiful

16.

German	**English**
a. Ein Dieb stahl unser Geld	A thief stole your money
b. Olu kommt	Olu is coming
c. Ich sah das Geld auf dem Boden	I saw it on the ground
d. Wie haben Sie das Geld ausgegehen?	How did you spend the money?

As observed in example (14) above, translating from English to either Yorùbá or German produces meaningful translations except that the translation of (14d) to Yoruba is devoid of accuracy and fluency in the target language. Therefore, it is easy to conclude that translating from English to either Yoruba or German may produce meaningful translation but this cannot be said of Yoruba to English translation as example (15) shows. None of the translations in (15) convey the meaning of the source sentence. This proves how difficult it is to translate from Yoruba to English. However, translation from German to English produces acceptable sentences as observed in (16).

It is therefore, plausible to conclude that besides the fact that there are more training corpus for English-German MT compared to Yoruba-English MT which has limited training corpus. English and German as languages have so much in common, they both belong to the West Germanic branch of the Indo-European language family. Yoruba, on the other hand, belongs to Yoruboid, a branch of the West Benue-Congo languages (Williamson and Blench 2000) which may not necessarily have much features in common with English. This, in our view will influence training, pruning, tuning and translation of any SMT. Consider example (14d) repeated as example (17) for convenience:

18. Let us work so that we would have money => Jẹ ki a ṣiṣẹ ki a yoo ni owo

It is observed that (17) is a complex sentence with independent and subordinate clauses. While it is plausible to accept that the independent clause is well translated in that it is an imperative clause with appropriate translation of both words and structures of the clause. But that could not be said of the subordinate clause. The subordinate clause is a mere substitution of words. "We" (first person subject plural pronoun) is substituted with its equivalence in Yoruba, *a,* e *would* is considered as a future tense marker of Yoruba *yóó* and the verb, *have*, is substituted for its equivalence, *ní* and same goes for translation of the Noun Phrase, "money", as *owó*. However, the equivalence of (17) should be something like *Jẹ́ kí á ṣiṣẹ́ kí á lè ní owó (lọ́wọ́)*. Consequently, the mis-translation of the auxiliary *would,* as future tense *yóó,* instead of its translation as a pre-verb adverbial, *lè,* brought about the loss of meaning in (17).

CHAPTER THREE

METHODOOGY

3.0 Preamble

This chapter discusses methods adopted to carry out the research. The discussion starts with the research design, instrumetation, data collection, data analysis procedure, as well as remarks and conclusion.

3.1 Research Design

This research is both exploratory and descriptive in outlook, therefore, methodological triangulation is adopted as its research approach. According to Karim (2007:3), methodological triagulation is a way of using both qualitative and quatitative methodologies in a research in order to get reliable findings. Qualitative research is geared toward having an in-depth understanding of a phenomenal and in this case, we intend to look at the translations produced by machine (the computer) in relation with the input data to assist in the explanation of well-formedness or otherwise of such translations. These translations will be compared with human translations quantitatively to ascertain the extent to which the computer can manipulate human language in line with the research questions raised in chapter one.

3.2 Instrumentation

The major instrument used is Moses. According to Moses User's Manual (http://www.statmt.org/moses/manual/manual.pdf), it is a Statistical Machine Translation (SMT) tool which has been employed and deployed by online translation systems like Google and Microsoft. The reason is that it is an open source language toolkit, flexible in terms of adaptability to any language pair. Moses has two main components: training pipeline and decoder. The training pipeline is really a collection of tools (mainly written in perl, with some in C++) which takes the raw data (parallel or monolingual corpus) and turn it into a machine translation model. The decoder is a single C++ application which, given a trained machine translation model and a source sentences, translates the source sentence into the target language.

Moses can be run on Windows through Cygwin. Cygwin is a large collection of GNU and Open Source tools which provide functionality similar to Linux distribution on Windows. It is also a DLL which provides substantial POSIX API functionality. As

Moses can be installed and complied on Linux and Mac system so can it be intalled on Windows but this is done under Cygwin. Therefore, with respect to this research, Cygwin was installed on Windows before Moses and other component software were installed. Some of the softwares are world alignment tools such as: giza++; language model tools: IRSTLM as well as other packages as obtainable in the Moses manual.

A computer laptop with 1TB 5400 rpm Hard Drive, 4th generation Intel(R) Core (TM) i7-4700 MQ Processor, 16GB DDR3 System Memory (2 Dimm) was procured for the purpose of this exercise. Other instruments used for preparing the data are: notepad++ and AntConc 3.4.4w 2014.

3.3 Data Collection

There are two kinds of data in this work: the input data and the output data.

3.3.1 Input Data

As it has been mentioned earlier that SMT uses equivalent translated texts as input data. Wikipedia seems to be the first point of call when free sources of this kind is needed but it was observed that the material on Wikipedia does not have the Engilsh equivalent. Besides, the source is written in the old orthography that does not depict the currect realities of the writing system of the language. It must be reported that there are some pages that have been impoved but they are very few. Another source of free material is the Jehoval Witness website where few equivalent translated materials are availble. For this research, we opted for literary translated texts because we believe these texts represent general domain to some extent. In that wise, the input data are purposefully selected. The table below shows the selected data:

Table 1:

S/N	Title	Author	Translated Equivalent	Translator
1	Ògbójú Ọdẹ nínú Igbó Irúnmọlẹ̀	D.O Fagunwa	The Forest of a Thousand Demon	Wọlé Sóyínká
2	Igbó Olódùmarè	D.O Fagunwa	In the Forest of God	Wọlé Sóyínká
3	Igbó Olódùmarè	D,O Fagunwa	The Forest of God	Gabriel A. Ajadi
4	Ìrìnkèrindò nínú Igbó Elégbèje	D.O Fagunwa	Adventure to the Mountain of Thought	Dapo Adeniyi
5	Àdììtú Olódùmarè	D.O Fagunwa	The Mysteries of God	Olu Obafẹ́mi
6	Aké: The Years of Childhood	Wole Soyinka	Aké: Ní ìgbà Èwe	Akinwumi Isola

3.3.1.1 Treatment of the Input Data

All the Fagunwa's books were written in the old orthography; in order that they conform to other materials and the modern realities of the writing system of the language, they were converted to the modern orthography so that the output will have the modern writing style of the language. We also had to break the novels done into equivalent sentences so that the number of sentences in a Yoruba novel is the same as the number of sentences in its English translated copy.

3.3.1.2 Error Index

In the first attempt to perform the experiment, while running the cleaning command, below was the error given:

```
clean-corpus.perl: processing /home/mrodoje/corpus/project-v7.yr-en.true.yr & .en to /home/mrodoje/corpus/project-v7.yr-en.clean, cutoff 1-80
/home/mrodoje/corpus/project-v7.yr-en.true.en is too short! at /home/mrodoje/mosesdecoder/scripts/training/clean-corpus-n.perl line 90, <E> line 922.
```

Checking manually, we found out that a sentence in Yoruba could be two or three sentences in English (vice versa). This does not allow for sentence alignment so as to have equal equivalent sentences. Consider example (1) below:

1a.

> *Èmi náà jókòó lé àga kan mo kọjú sí i. Bí mo ti ń jókòó ní ó bẹ̀ mí pé ki n fún òun ní omi mu,* (2005c:2)
> *I also chose one for myself and proceeded to place my buttocks on it when he begged the favour of a drink.*

1b.

> *Nípa ti ìyàwó mi, mo lè wí fún ọ pé ìyàwó mi ń bẹ; nípa ti àwọn ènìyàn mi, mo lè wí fún ọ pé alàáfíà ni àwọn ènìyàn mi wà; bí ó sì tilẹ̀ jẹ́ pé, ọjọ́ pẹ́ tí mo ti fi àwọn ènìyàn mi wọ̀nyí sílẹ̀, síbẹ̀ àdúrà tí ọmọ ènìyàn bá gbà fún ara rẹ̀ ni Olódùmarè ń bá a gbà, ọkàn ẹni ni àlúfáà ẹni* (2005b:78)
>
> *As for my wife, I am able to assure you that I do have a wife. My people, I also assure you, are all in good health. Even though it has been a while since I left them, however, it is the prayer that the son of man prays for himself that Edumare grants, one's heart is ultimately one's priest.*

You will observe that in (1a) the English version is a sentence while the Yoruba equivalent is two sentences. In (1b), the Yoruba sentence is just a sentence represented in three sentences in its English equivalent translation. Even though the translations portray the meaning of both the source and target sentences; automatic sentence aligment will not properly marge the sentences. To an automatic sentence alignment, punctuations like fullstop, exclamation mark, question mark etc mark the end of a sentence. This may lead to a mis-merge which may consequently lead to mis-translation.

Another significant observation is that some sentences in the source text were completely ignored. Probably, the translator found them insignificant or it could be an error of ominssion on the part of the translator. This cuts across all the literary text that are used. Only a few of them are examplified below. In another instances, some elements were introduced in the translated text that were not in the source text This could be an instance of ingenuity of the translator but they could affect word-word alignment. All these we needed to manually correct before we could build our input data.

Consider (2), an extract from *Ake: The Years of Childhood* below:

2a. Source sentence: 'Don't mind her' **I told Osiki.**
Target sentence: Má dá a lóhùn ọjàre.
2b. Source sentence: I heard the confused boy calling on God to save him from the stigma of becoming a murderer in his lifetime.
Target sentence: Mo gbọ́ tí ọmọ ọ̀hún ń bẹ Ọlọ́run pé kí òun má mà ní ẹ̀jẹ̀ ènìyàn lọ́rùn layé òun o. **Ẹ̀rù ti bà á**.

The bolded sentences are the missing and the added part in (2) above. While (2a) is an instance of omission, (2b) is an instance of ingenuity. Be it as it may, both will not result into perfect merge alignment which computer will derive its probalility from.

Based on the aforementioned, it is difficult to do automatic sentence alignment with the available software (http://mokk.bme.hu/en/resources/hunalign/). Therefore, we have to devise a means of ensuring that the instances selected in the data have equal sentences in both English and Yoruba corpora manually before automatic sentence alignment.

Another noticeable error is wrong translations. For example, in *Ìrìnkèrindò, ọmọ,* which is known to be equivalent to *child* was translated as *friend* and no context suggests or prompts this out of context translation as the sentence (3) below proofs:

3.

Yorùbá sentence: *Yára tètè lọ kí o pè é kí o wá rí mi, ọọku o, ọmọ mi.*

English translation: *Go on now and bring this man to me. I greet your energy my friend'.*

As a result of this, the computer will assign a weighted score for *ọmọ* as a possible translation of *friend.* This may affect the translation of *ọmọ* as *child* depending on the frequency of occurence of such a translation in the training corpus. And by implication, there is the possibility that computer produces *ọmọ* as an equivalent translation of *friend*. This, to an end user, would be regarded such error would be attributed to the computer without realizing that the error is actually from the training corpus as observed above. Other errors include typographical errors, printing repetition, wrong arrangement etc. All these errors constitute the error index table below based on each input material.

Table 2: Error Index

			Types and percentage of the error									
	Text	**Translator**	**Translations not in the source text**	**% of Error**	**Source sentence not translated**	**% of Error**	**Error of Translation**	**% of error**	**Printing and publication error**	**% of error**	**Total number of lines**	**Total**
1.	*Igbó Olódùmarè*	Wole Soyinka	10	0.44	36	1.58	4	0.17	2	0.09	2280	2.28
2.	*Igbó Olódùmarè*	Gabriel Ajadi	6	0.27	29	1.29	12	0.53	144	6.38	2256	8.47
3	*Irinkerindo*		21	0.86	28	1.17	19	0.79	-	-	2400	2.82
4	*Ogbójú Ọdẹ níní Igbó Irúnmọlẹ̀*	Wole Soyinka	3	0.17	6	0.34	2	0.11	1	0.06	1754	0.68
5	Aké	Akin Isola	3	0.25	4	0.33	2	0.16	1	0.08	1221	0.82
6	Yoruba: Intermediate Text		-	-	-	-	-	-	-	-		
Total			**43**	**1.99**	**103**	**4.71**	**39**	**1.76**	**148**	**6.61**	**10119**	**15.07 %**

The overall percentage error (3.29%) may look insignificant but the input data could be improved if these errors are corrected before training. Then one is sure that other errors generated after the training would be attributed to the computer and the procedural pipeline processes.

3.3.2 Output Data

Since our interest is not just to build a translation machine but to evaluate its output whatever the computer can generate as its translation constitutes data for us. Firstly, we would conduct human evaluation like we did for translations from Google Translate in section 2.8 using Akungba and Ibadan Structured Sentence Paradigm. Then we would examine output sentences in line with the research questions raised in chapter one.

3.4 Data Analysis Procedure

There are two kinds of analysis we intend to do. First is quantitative analysis, measuring machine translation in relation with human translation. Akungba and Ibadan Structured Sentence Paradigm were used as tools to elicit machine translation outputs which are compiled as questionnaires for human judgement. The human and machine translations are then subjected to analysis using SPSS. The second analysis which is qualitative would examine the translation of machine in line with the research questions raised in chapter one.

3.5 Moses Pipeline and Procedure

A pipeline is the continuous and somewhat overlapped movement of instruction to the processor or the arithmetic steps taken by the processor to perform an instruction (see whatis.com). Appendix 2 is the complete Moses pipeline used for this exercise, but in this subsection, we intend to show some results of the training processes.

3.5.1 Tokenization

Before training, there is the need to prepare the raw data and the first is tokenization. According to Koehn (2010: 34), tokenization is breaking up of a raw text into sentences which are strings of words and punctuations which are to be separated by space. The command line to achieve this is found in the appendix as stated above. Since Yoruba and English more or less have similar writing systems compared with languages like Japanese or Chinese that use logographic system, the process was achieved easily.

3.5.2 Truecasing

Truecasing in translation models is a process of dropping distinctions between uppercase and lowercase, be it at the beginning of a sentence (The), in the middle (the) or all-caps heading (The). Lita, Ittycheriah, Roukos and Kambhatla (2003) states that truecasing enhances the quality of case-carrying data, brings into the picture new corpora originally considered too noisy for various NLP tasks and performs case normalization across styles, sources, and genres. The command line is in appendix 2.

3.5.3 Cleaning

It is a process of removing mis-aligned, long, as well as empty sentences as they can cause problems with the training pipeline (Koehn 2015:36). The cleaning was limited to 80 words. This reduces our corpus from 16,146 sentences to 14,673 sentences.

3.5.4 Language Model Training

According to Koehn (2010:9) Language Model (LM) measures the fluency of the output and is an essential part of Statistical Machine Translation (SMT). Some of the software, used for language model estimation are IRSTLM and KenLM. KenLM is adopted for this work because it is included in Moses as default in the Moses tool-chain. We also base our n-gram language model with n=3. Running LM command in the *LM Direcory*, below is the result:

```
Chain sizes: 1:208032 2:2139856 3:5558420
=== 5/5 Writing ARPA model ===
Name:lmplz      VmRSS:5620 kB    RSSMax:5628 kB   user:10.531      sys:14.562
CPU:25.093      real:41.677
```

This we need to binarize the .arpa.en file using KenLM for fast loading. Then we need to query our LM with a Yorùbá sentence: *Adé fẹ́ràn owó.* Below is the result produced after the query:

```
Adé=6384 1 -5.77036     fẹ́ràn=141 1 -3.27466    owó=258 1 -3.39753      </s>=2 1 -
2.33102      Total: -14.7736 OOV: 0
Perplexity including OOVs:     4936.19
Perplexity excluding OOVs:     4936.19
OOVs:   0
Tokens: 4
```

```
Name:query	VmRSS:4256 kB	RSSMax:4284 kB	user:0.015	sys:0.015	CPU:0.03
real:0.0233152
```

The result shows how likely each of the words in the sentence will follow each other from the available corpus.

3.5.5 Training the Translation System

After the LM, we need to train the translation system. We need to run word-alignment using GIZA++ to generate phrase extraction and scoring, create lexicalised reordering tables, and then create Moses file configuration with a single command. The training took 1.31 hour. At the end of the training, *moses.ini* file was created at the directory ~Working/train/model. The ini file could be used to decode (that is, to translate) although according to the moses manual, there could be a couple of problems with that. The first problem is that it's very slow to load hence, we need to binarise the phrase table and reordering table. This means that we have to convert them to a format that can be easily loaded. Another problem is that the weights used by Moses to waight the different models against each other are not optimised. Moses.ini file is set to default values of 0.2, 0.3. To find better weights, we need to tune the translation system which is the next step.

3.5.6 Tuning the translation system

To do the tuning, we need another set of corpora. This may not be as big as the training corpus and should be separate from the training corpus. This corpus is in the corpus file. We equally need to prepare the corpus for the tuning training by tokenising and truecasing it. We could do the tuning process with this command at the Working directory:

```
nohup nice ~/moses/scripts/training/mert-moses.pl \
~/corpus/EnglishYorubaTuning.yor-en.true.yor	~/corpus/EnglishYorubaTuning.yor-en.true.en \
~/moses/bin/moses train/model/moses.ini --mertdir ~/moses/bin/ \
&> mert.out &
```

The process took several hours. The tuning started at April 23, 2016, 3:27:30PM and ended at April 24, 2016, 2:59:59 AM. The end result of tuning is an ini file with trained weights which is in ~/working/merit-work/moses.ini.

3.5.7 Test

This allows us to test translation with some sentences. The command used is:

```
~/moses/bin/moses -f ~/working/mert-work/moses.ini
```

With the command, it is possible to test ‘moses’translation. Although, it is also possible to binarise the phrase table and lexicalised reordering models for the decoder to start up quickly. To do this, we need to creat a directory and then binarise. The command is in the appendix. With the binarization, loading the decoder and translation bacome faster. The translation at this point may not be good enough hence we could to metrics evaluation. To do this, we need another corpus. For this purpose, we generate 180 sentences from Jehovah Witness website. The corpus was preprocess by tokenising and truecasing it. Thereafter, re filtered our trained model by removing entries that are not needed for translation test so as to make translation faster. Then we test the decoder by first translating the test set before running the bleu script. At the end of running the bleu script, below is the result:

```
~/moses/scripts/generic/multi-bleu.perl \
> -lc ~/corpus/bleutest.true.en \
> < ~/working/bleutest.translated.en
BLEU = 2.51, 25.7/4.5/1.1/0.3 (BP=1.000, ratio=1.347, hyp_len=4064, ref_len=3018)
```

With the result of Bleu test, we have come to the end of baseline system of moses as a Statistical Machine Translation toolkit.

CHAPTER FOUR

PRESENTATION AND ANALYSIS

4.0 Preamble

This chapter presents linguistic data from the translations of our moses SMT and the analysis is done in line with the research questions raised in chapter one. Therefore, the analysis does not explore any translation theory rather descriptive analysis is the bedrock of our presentation in this chapter.

4.1 Computer Acquisition of Natural Human Language

The fact that little children acquire a complex system (language) with little or no effort while with all sophisticated methods or approaches to language modelling, the process is still a challenge to scholars and artificial intelligence experts till date remains a puzzle (Manaris 1998; Bonache and Jiménez-Lopez 2011; Wintner 2010; Briscoe 2013). Some have argued that language remains an inherent property of human intelligence. As good as the proposition of innatist may be, there is still a need for the environment to activate the innateness (that is, the language of the immediate environment). This we have argued in chapter two. Wintner (2010:86-88) opines that language learning tasks (whether human or computer) rely on the existence of large text corpora that document language use, both for training and for evaluation. He further states that Computational linguistics tasks standardly use manually-annotated sentences from the Penn Tree Bank (PTB). To him, computational approaches to language learning from data can be distinguished along the axes below:

- Data
- Task
- Grammar
- Evaluation

4.1.1 Data

Data serves as input to language learning[13]. As mentioned above, there is a need for large text corpora for language training. For example, Denkowski, Hanneman and

[13] Language acquisition and language learning are used interchangeably in the literature however, computer learns natural human language through language modelling while a child acquires a language based on the language of the immediate environment.

Lavie (2012:261) cite Koehn (2005) to have got parallel corpora for English-French MT from European Parliament proceedings, United Nation document and News Commentary totaling 13 million sentences. They then built their MT with 27 million sentences given the detail of their data below:

Corpus	Sentences
Europarl	1,857,436
News commentary	130,193
UN doc	11,684,454
Giga-FrEn 1stdev	7,535,699
Giga-FrEn 2stdev	5,801,759
Total	**27,009,541**

Source: Denkowski, Hanneman and Lavie (2012)

It is from the corpora that the computer learns natural language. Generally, it is argued that the more the corpus the better the training output (Koehn 2010, Lavie 2012). In our case, 16, 146 sentences is the size of our parallel corpora:

Corpus	Sentence
Igbó Olódùmarè (Soyinka)	2, 280
Igbó Olódùmarè (Ajadi)	2, 280
Irinkerindo	2, 400
Ogbójú Ọdẹ níní Igbó Irúnmọlẹ̀	2, 256
Adiitu Olodumare	3, 362
Aké	1, 638
Bible (Genesis and Exodus)	1, 930
Total	**16, 146**

Our parallel corpus is too small, if it is compared with that of Denkowski, Hanneman and Lavie (2012) data above.

It should be noted that our training and evaluation corpora are not annotated. Hence, the computer heuristically tagged Yoruba words. This greatly affect Yoruba lexical entries and the weight assigned them. Homophones are seen as one and single words without considering their syntactic and semantic functions.

For example the word *bí* could be a verb (to give birth), it could be Yes/No question marker, an adverb as well as a complementizer as expressed in the example (1) bellow:

1a. Toyìn bí Fẹ́mi
Toyin born Fẹ́mi
'Toyin gave birth to Femi'

b Toyìn jẹun bí?
Toyin eat Q
'Did Toyin eat?'

c Bí mo bá ní owó màá kọ́ ilé
Com 1st adv have money will build house
If I have money, I will build a house

d Nígbà tí mo ti mọ bí ọmọ ọdún mẹ́wàá ni mo ti ń bá bàbá mi lọ oko ọdẹ
PP rel1st asp reach adv child year ten foc 1st asp prog adv father me go farm hunter
‘Since I was like a child of age ten, that was when I havw been following my father to game’

e Gẹ́gẹ́ bí ìwà ẹdá yín pẹ̀lú, ọkàn yín kìí balẹ,
Just adv behaviour human you with, mind you does not settle
‘Like you behaviour, you do not have settled mind’

In (1a) above, *bí* functions as verb therefore its semantic content is different from other *bí*s in the other examples. However, *bí* in (1b) is a Yes/No question maker which is adjoin to simple declarative sentence final position. While *bí* in (1c) a complementizer and *bí* in (1d and e) are preverb adverb. *Bí* as complemetizer was given more wieght because of its frequencey of occurence in our corpus coupled with the fact that annotation or in other words part of speech tagging to show this kinds of difference was not done to our training corpus. Our system sees *bí* as a lexical entry which explains the error in translation of a sentence like (2) below:

2.	**Source sentence**	**Translation by our computer**
	Tóyìn bí Fẹ́mi	Tóyìn if Fẹ́mi

To overcome a challenge like this, we suggest that languages with limited resources like Yorùbá should annotate their corpora as to avoid mistranslation like (2) which can be rectified by a simple part of speech tagging.

Summarily, Yoruba-English statistical machine translation is faced with two challenges as regard data: there is need to create more data and annotate them.

4.1.2 Task

Wintner (2010:88 and 95) explains the term *task* as what a learner is required to learn which could be learning a language as a set of string or these strings as augmented by structures. He suggests that computational models of language acquisition should focus on the easier task of learning language as a set of strings, leaving the induction of syntactic structures to future research.

Our model is to learn Yoruba and English as a set of string for the purpose of unidirectional translation. Thus, can we then say our model as acquired enough string for the unidirectional Yoruba-English translation. This question will appropriately be answered in the subsections below.

4.1.3 Grammar

Grammar in this sense is formal grammar not necessarily linguistic grammar. Wintner (2010:88) explains that grammar induction algorithms are usually formal and explicit in their definition of the class of models that they attempt to learn. These can be deterministic finite-state automata (FSA) or Hidden Markov Models (HMMs) or Probabilistic context-free grammars (PCFGs) or Tree substitution grammars, with a variety of probabilities models. In our case we adopt Hidden Markov Models.

Ghahramani (2001:2) explains that a hidden Markov model is a tool for representing probability distribution over sequences of observation. The observation could then be denoted at time *t* by the variable Y_t. He further states that this can be a discrete alphabet, a real-valued variable, an integer, or any other object, as long as we can define a probability distribution over it. It could also be assumed that the observations are sampled at discrete, equally-spaced time intervals, so *t* can be an integer-valued time index.

4.1.4 Evaluation

Following Wintner's (2010:88) explanation, grammar induction algorithms are evaluated on annotated data; it is expected to learn the bracketing and sometimes labels in the corpora that are manually annotated. In our case, we use manual or human evaluation instead of automatic evaluation. It is assumed that SMT should be trained and evaluated. ' As elucidated in chapter two, where we have the average mean score of Google Translate of translating Yoruba language to English language for fluency is 1.449 while that of adequacy is 1.362. When we compare this result with that of our computer translation, they are more or less the same just that our computer had a little higher score than google translate. For fluency it was rate 2.192 while that of adequacy 1.972.

Approximating to the next decimal point, it is noted that both Google Translate and our machine in terms of fluency are rated 2.0. Meanwhile, our computer is rated higher when raw score is compared. The same goes for that of the adequacy. Comparing raw score both are rated 1; if we have to approximate, our machine will be rated 2 while Google Translate will still be rated 1. At this, there will be a clear difference in the sense that Koehn's (2010:232) table on adequacy equate 1 to NONE meaning that the translation is not producing meaningful translation while 2 is equated to little meaning. In other words, but Google Translate and our machine produce translations of little meaning because they are both rated 2. On the contrary, our machine is rate two in terms of fluency which means Disfluent compared with Google Translate which is rated 1; meaning that it produces translation that are incomprehensible.

The rating as stipulated above actually show in the table1 below:

Table 1:

	Yoruba	**Google Translate**	**Ibadan SMT**	**Human**
1	kí ni o tún ń dúró dè?	What are you waiting for?	What did you also disgorgement?	What else are you waiting for?
2	kò wá sí ìpàdé	not attend	Not come to the meeting	He did not come to the meeting
3	èwo ni o rà?	Which is bought?	Èwo was he bought?	Which one did you buy
4	Mélòó ni o rà?	Several have bought?	How was he bought?	How many did you buy?
5	kí ni orúkọ rẹ?	What is your name?	What is your name?	What is your name?
6	Ta ni bàbá rẹ?	Who is your father?	Who is your father?	Who is your father?
7	Olú rí mi	I saw capital	Olú saw me	Olú saw mi
8	Olú rí ẹ	Headquartered see you	Olú saw	Olú saw you
9	Olú rí ọ	Headquartered see you	Olú see you	Olú saw you
10	Olú rí i	Realized capital	Olú saw	Olú saw him/her
11	Olú rí yín	Your capital	Olú saw you	Olú saw you
12	Olú rí wọn	Their capital	Olú saw them	Olú saw them
13	Mo rí Olú	I saw Alejandro	I saw Olú	I saw Olú
14	O rí Olú	You see Emperor	You saw Olú	You saw Olú
15	Ó rí Olú	He saw Emperor	He saw Olú	He saw Olú
16	A rí Olú	We found Alejandro	Ah saw Olú	We saw Olú
17	Ẹ rí Olú	See Emperor	Ẹ saw Olú	You saw Olú
18	ẹ rí Olú	See Emperor	You saw Olú	You saw Olú
19	Ìwọ rí Olú	You see Emperor	You saw Olú	You saw Olú

20	Òun náa rí Olú	He saw Emperor	FARINA saw Olú/Olú the FARINA	He also saw Olú

From table (1) only data serial number (5&6) are correctly translated by the tree mechanism—google translate, Ibadan SMT and Human in term of adequacy and fluency. Like it was pointed out in 2.5.8, Google Translate had difficulty in translating *Olú* with the notion of refering to human's name. This Ibadan SMT did correctly. Although if *Olú* is typed with lower case or it is typed alone, the machine translates it as *mushroom* (which is another interpretation if olu is not referring to a human) otherwise *Olú* remains without being translated.

Another commonality of both Google Translate and Ibadan SMT is their errors. Serial number (2,3,8,10,16,20) show that both machine translated the sentences wrongly. It is observe that serial number (2) from table (1) is an indication of difference in the language structures. While the High Tone Syllable marking the third person singular subject pronoun in the Yoruba negative sentence is always covert unlike its declarative counterpart which is overt, this we expect in the translation but the machines took the source sentence living out the subject pronoun.

Serial number (3) from table (1) is an instance of incomplete thought. While there is an iota of meaning denoting something being bought yet the translation did not convey complete sense in its English equivalent. Serial number (4) is an example of mistranslation. Its English equivalent is a total deviation from the source language sense. Consequently, serial number (16 and 20) are not just mistranslated example but also an instance of introducing terms alien to the source sentence. Both *Alejandro* and *Ah* are alien to example (16) in relation to the source sentence while *FARINA* is alien to example (20). We suspect that *A* and *Ah* are used in the corpus as equivalence in the training corpus hence the machine may not recognise *A* as a pronoun in a Yoruba sentence. That being said we are still puzzled how *Alejandro* and *FARINA* are used in the examples.

Serial number (8 and 9) in the table 1 above shows how Ibadan SMT handles Yoruba second person singular object pronoun. While ọ is translated correctly, its ẹ equivalent is dropped in the translation. This means that whenever ẹ is used, there will not be any translation for it unless its ọ counterpart is being used. The reason for this error is found in the training corpus. In the training corpus, it is only ọ that was used as the second person singular object pronoun which makes it difficult for the machine to recognise ẹ as pronoun in the object position. Third person singular object pronoun which is lengthen of the last syllable of the verb is also not translated because there is not much weight assigned to them for proper translation (see serial number 10).

However, Ibadan SMT translates Yoruba subject pronouns correctly except Ẹ which is the secon person plural subject pronoun. The machine makes a demacation between the upper case and lower case of Ẹ. Upper case Ẹ is not translates but whenever lower case (ẹ) is used then the machine translates it correctly as *you* with its plural sense (serial number 17 and 18).

At this point we can attempt to answer one of the research questions in section 1.8. If it is possible for machine to make errors in translation the so called major languages with huge corpora as well as languages that are more related in terms of language family and structure like German and French as against English how much more will it make error to a Yoruba language that is distance to English in terms of language family and structure. Therefore, it is plausible to conclude that language acquisition is restricted to humans hence machine could only model human language.

Language is born out of exposure and experience. This postulation is neither behaviourist not innate position rather a common grand for both theorist. Both theorist are of the opinion that environment is important to language acquisition. While behaviourists believe language is learnt from the environment, the innatists believe that language is biological but needed exposure to the environment. For a computer machine or robot the only environment available is the corpora to learn from. The question to ask is, is it possible and plausible to reduce all words and sentences of a language to a text no matter how large it is? In our opinion, language is innovative which relies on human cognition and reasoning which is difficult for computer to achieve hence, human being can acquire a language while a computer machine can model a language.

Furthermore, we should say that machine can model a human language to near acquisition if and only if human beings will provide all necessary information that will assist a machine to learn from, including pragmatic information. Therefore, this goes beyond disambiguation in terms of structure and meaning but also culture peculiar information.

4.2 Literary Text and SMT Development

To build a SMT, there is a need to have training corpus which could be bilingual or monolingual. Most often, researchers prefer bilingual corpora. Because there are available open source sentence alignment tools especially for the “major” languages. For example, Geometric Mapping Alignment by Ali Argyle, Luke Shen, Svetlana Stenchikowa and I.Den Melamed who design the tool for languages like French/English, Malay/English, Romanian/English, and Russian/English. Another alignment tool is Hunalign-Sentence Aligner. It was developed under the Hunglish project to build the Hunglish corpus. Hunalign aligns bilingual text on the sentence level. Its input is tokenized and sentence-segmented text in the two languages while its output is a sequence of bilingual sentence pairs. Like most sentence aligners, hunalign does not deal with changes of sentence order: it is unable to come up with crossing alignments, i.e., segments A and B in one language corresponding to segments B’ A’ in the other language. Hunalign was written in portable C++. It can be built under basically any kind of operating system.

As good as automatic sentence aligners are, they are often used for technical translated materials which are straight forward in their translations. It would be difficult for automatic alignment to appropriately align translated fictions. This is in line with Xu, Max and Yvon (2015:5) submission that literary text should be more difficult to align than say, technical documents. This submission of theirs is informed by Langlais and Véronis (2000) unsatisfactory results achieved by all sentence alignment system during Arcade Evaluation Campaign Xu, Max and Yvon (2015:5) quoting Langlais et al (1998) to have said:

> These poor results are linked to the literary nature of the corpus, where translation is freer and more interpretatives.

This is one of the challenges encounted in the building a bilingual text for Ibadan SMT corpus. As pointed out in section 3.3.1.2 a sentence in a literary material may be two or three sentences in its translated equivalence. For convince cosider the example (3) below:

3 a Kò sí bí a ti le ṣe ifa kí ó má hùwà èkùrọ́, kò sí bí a ti le ṣe ẹni ti o ń káàkiri ẹ̀yìn odi kí ọgbọ́n rẹ̀ má ta ẹni tí kò kúrò lójú kan yọ.

b But is it really possible to act Ifa, the oracle of divination in a stage-play without one's behaviour coming close to that of the palm-nut? The man who travels countrywide must embody within himself a whole load of knowledge and understanding which surpasses that of one who restricts himself to his own house.

c Even the open lawns and broad paths, bordered with whitewashed stones, lilies and lemon grass clumps, changed nature from season to season, from weekday to Sunday and between noon and nightfall.

d Kódà, àwọn pápá ìṣeré àti àwọn ọ̀nà fífẹ̀ tí a to òkúta tí a kùn lẹ́fun sí lẹ́gbẹ̀ẹ̀gbẹ́, àti àwọ lílì àti ṣiiri lemon grass máa ń yí ìrísi wọn padà láti ìgbà dé ìgbà. Bí wọn ṣe rí lọ́jọ́ lásán yàtọ̀ sí ti ọjọ́ ìsimi, bí wọn sì ṣe rì lọ́sàn-án yàtọ̀ sí bí wọn ṣe rí lálẹ́.

It is observed that (3b) is different from (3a) in form and function. By form it means the sentence structure and by function, meaning. (3a) is a compound complex declarative sentence. (3b) is the translation of (3a) and one will expect that both should be the same in meaning and structure. But unlike (3a) which has one sentence, (3b) has two sentences. The first sentence of (3b) is a complex interrogative sentence. While the second sentence of (3b) is a compound complex declarative sentence. It is evident that there is nothing interrogative in (3a) as well as nothing connoting stage-play in it. It could then be concluded that the insertion of the stage-play in (3b) is the interpretation of the translator as observed by Xu, Max and Yvon (2015:5) quoting Langlais et al (1998). Equally (3c) is a compound complex sentence whereas, its equivalent (21d) are two sentences. In other word, automatic aligner will regard (3c) as a string will (3d) as two strings. Then one of the (3d) will be mis-aligned or deleted in the training process. Therefore, these king of differences affect automatic alignment and training of a statistical machine translation hence manual alignment was adopted for our training corpus which is time consuming and energy tasking.

Literary texts for a resource scarce language like Yoruba play a significant role in developing a SMT. Like it was pointed out in section 1.1 that to develop a SMT there is a need for large volume of translated text. Technical translation would have been preferred because of its direct translation which machine could easily learn from with minimum re-ordering rule to conform to the structures of the language pairs. For instance, consider example (4) below from some technical materials:

4a i. Ó ga tó mi
HTS tall as me
's/he/it is as tall as me'

ii. Ó ga tó
HTS tall enough
's/he/it is tall enough'

iii. Ó kéré jù
HTS small much
's/he/it is too small' (adapted from Schleicher (1998:17))

4b i. Lọ ya méjì
Go tear two
'Go tear (off) two'

ii. Mo fàwé mi ya
1sg pull-book 1sg apart
'I tore my book'

iii. a gba ọ̀kan
3pl recieve one
'We got one' (adapted from Stevick and Arẹ̀mú (1963:17))

4c i. Adé bẹ olùkọ́ ní iṣẹ́
Adé beg teacher prep work
'Ade sent the teacher on an errand'

ii. Òfófó pa aláròká
Tell-tale kill gossip
'Tale-bearing killed the gossip'

iii. Iṣẹ́ ajé sọ ọmọ nù bí òkò
Work business fling child away like stone
'Daily bread has flung son of man far-afield'
(adapted from Yusuf (2011:262))

Example (4a-c) are extract from teaching materials. While (4a&b) are proficiency based (4c) is meant to teach Yoruba sentence structure. Although, there are obvious differences in the structures of the source language and its equivalence in the target language like insertion of the copular in example (4a); differences in the structure of Noun Phrase like the need for article in the English equivalence which is not necessary in the source language as found in example (4ci&ii) as well as Noun, Determiner juxtaposition as found in example (4bii). Examples (4) are all simple declarative sentences except example (4bi) which is a simple imperative sentence. If (4) is compared with (5) bellow:

5a. Ayé kún fún ibùgbé ìyanu, Ọba bí Ọlọ́run kò sí.
World full for dwelling-place miracle, king like God ned is
‘Marvels fill the world over, there is no king as God.’

b. But the objects on which my eyes were fastened were two black
heavy-snouted tubes mounted on wooden wheels.
(Soyinka 2000)

‘Àwọn àgbà dúdú méjì kan onímú ẹlẹ́dẹ̀, tí wọn gbé lérí kinní
kan tí ó jọ ọmọlanke, ni mo ń wò nítèmi ṣáá.’(Isola 2001)

c. Lọ́jọ́ kejì, o fi ẹrù iṣu méjìlá ránṣẹ́ pẹ̀lú ṣago ẹmu ogidi mẹ́wàá,
ó gbé àpò ìdọ̀họ gaàrí kan, àpó èlùbọ́ méjì, garawa epo mẹ́jọ pẹ̀lú
àpó iyọ̀ ńlá kan. Fagunwa (2005c)

‘He did not pause there, he sent us twelve sack-loads of yams
and pure frothing palm-wine, made us trample on gari, sacks of
yamflour, eight giant gallons of kernel oil and outsized sachets
of salt.’(Adeniyi 2000)

Examples (5) are compound and complex declarative sentences unlike (4) which are simple declarative and imperative sentences. It is also observed that (5) have complex word order. For example, the word *Ayé* appears as sentence initially position in example (5a) but appears word medially in its English equivalent. This will influence reordering rules in reflecting the structures of the language pairs. Example (5b) is a good translation of English to Yoruba even though it may be somehow difficult for a word to word machine to get equivalence of each of the words because the opening clause in the source sentence is ending the target sentence. More so, the Yoruba equivalence of (5b) in an interpretation of the translator trying describe the sources sentence in the cultural acceptable for of the target sentence. While this is acceptable for conventional translation there will be a need for proper annotation for machine to adequately master the strucues of the languages. (5c) is an example of mis-marge of structures. *Lọ́jọ́ kejì*, is never a direct equivalence of *He did not pause there.* As much as both are necessary for the narratives they could cause a mis-translation for the machine.

However, literary texts constitute 14, 216 out of 16, 146 which is 88% of the total data used. It is also needful to state that the texts are not annotated hence the computer heuristically classified Yoruba words for translation based on frequency of occurrence and suggested translation from the training corpus as could be observed in the table (2) below:

Table 2:

S/N	Source Sentence	Ibadan SMT	Human Translation
1	Òun náà rí Olú	He also saw Olú	He too saw Olú
2	Àwa pàápàá rí Olú	Àwa himself saw olú	We too saw Olú
3	Olú kò dé ibẹ̀ rí	Olú did not arrive ibẹ̀ saw	Olú had never been there before
4	Olè jí owó Òjó	Olè woke money Ojo	Thief stole Ojo's money
5	Olè jí owó wa	Olè woke our money	Thief stole our money
6	Olè jó owó wọn	Olè woke their money	Thief stole their money
7	Mo wọ aṣọ funfun	I wore white clothes	I wore white clothe
8	Olú kò wá rárá	Olú not come at all	Olú did not come at all

From table (2) above, serial number (1&6) could be accepted as translation of the source sentence but serial number (7) is does not conform to the structure of English language. In (2) the Long Pronoun is seen as a noun with anaphora *himself* to show the emphasis in the source sentence and this informs the mis-translation which people who do not speak Yoruba at all may not get the gist in the source sentence when the look at the translation. (3, 4&5) reflect the ambiguity in the Yoruba verb *jí*. While *jí* could mean 'to wake up', it could also mean 'to steal' in the same structural position. Consider example (6) below:

6 Wọ́n jí Ìyábọ̀

(6) could mean *they woke Ìyábọ̀ up* or *Ìyábọ̀ has been stolen.* However, *jí* occurs in 123 times in the training corpus. Out of this 123 occurances, *jí* in the sense of 'stealing' only occurs 17 times making 13.8% while *jí* in the sense of 'weaking up' occurs 106 times making 86.1%. Therefore, *jí* in the sense of 'to weak up' is assigned a higher weight score which necessitate the translation of *jí* as 'woke' in the table (2) above.

Another wrong translation in the table (2) is the use of the verb *rí* which the computer translated correctly as seen in serial number (1&2) but when *rí* is used as an adverb in serial number (3) the computer could not differentiate the verb and the adverb.

The computer had to translate the adverb as the verb which informs the mis-translation. For example *rí* occurs 2233 times in the training corpus and *rí* as adverb only appears 20 times hence the over generalisation that warrant (3) in the above table.

From the aforementioned, it is obvious that the available literary texts are not sufficient enough as a training corpus for a SMT. It is not enough to say they are not enough, effort should also be geared towards part of speech tagging in order to avoid ambiguity like it is observed in the table (2). Gavrila and Vertan (2011:555) suggested that it is not large amount of training corpus that is important but test size, type and evaluation procedure. In their words:

> … we can conclude that for technical domains a small, manually corrected corpus can be successful used for obtaining a reasonable translation output… all the results we have presented reinforce the idea that SMT is fully dependent on training and test size and type and on the evaluation procedure…

This implies that even if we have a very large corpus that are not appropriately corrected and we have procedural issues with training, test and evaluation, we may not necessarily get a reasonable output.

4.3 Moses and Tone

Tone in Yoruba like many other African languages is performing more than Phonological functions, it also performs syntactic and semantic functions. Consider example 7 below:

7a Adìẹ funfun
Hen white
White Hen

7b Adìẹ́ funfun
Hen White
The Hen is white

It is observed that the tone difference in the last syllable of the first word brings about the difference in structure and maning of (7 a&b). While (7a) is a Noun phrase, headed by the noun *Adìẹ* with its adjectival qualifier *funfun;* (7b) is a sentence. The subject of the sentence is *Adiẹ* as is predicate is the verb *funfun.* What brings about the diffrence in the structure and meaning is the tone change in the last syllable of *Adìẹ*. This is what was tested in Moses using Owolabi (2013:261) explanation to ensure whether Moses could recognise tone change in these structures for the purpose of translation.

4.3.1 Verbs with NP Objects

Owolabi (2013:261) explains that when a verb has low tone as its inherent tone, and the verb subcategories for an NP object, the verb losses its inherent low tone and assume mid tone. Each word was first tested to ensure whether Moses could identify the words in isolation. Therefore, words like *sùn*, *rà* and *tà* were translated as *slept*, *bought* and *sell* respectively. When each of these words take NP objects, the lose their inherent tone for mid tone as examples 8 below show:

8a Mo ta ìwé
1sl sell book
'I sold a book'

8b Mo ta bàtà
1sl sell shoe
'I sold a shoe'

8c Mo tà púpọ̀
1sl sell much
'I made much sales'

8d Mo tà púpọ̀ ní àná
1sl sell much prep yesterday
'I made much sales yesterday'

8e Mo sùn
1sl sleep
'I slept'

8f Mo sun orun
1sl sleep sleep
'I slept'

8g Mo sun ilé
1sl sleep house
'I sleep at home'

8h Mo sùn ní ilé
1sl sleep prep house
'I slept at home'

8i Mo sùn gan-an ni

1sl slep much foc
'I slept so much'

8j Mo ra aṣọ
1sl buy clothe
I bought a house

8k Ẹlẹ́mu mí tà
Palm-wine seller my sell
'My palm-wine seller made a good sales'

You will observe that verbs in example (8a, b, f, g and j) took direct NP object hence losing their inherent low tone for mid tone. This change in the tone affect the translation of those sentences by Moses. Example (8a, b, f, g and j) are repeated below with their Moses translation as (9 a-d).

9a Mo ta ìwé => I who book

9b Mo sun orun => I roasted sleep

9c Mo sun ilé => I roasted home

9d Mo ra aṣọ => I bought clothes

It is observe that only (9d) is translated by Moses correctly probably because it is not ambiguous in its usage unlike (9a-c) while *ta* could be mistaking for the nominal interrogator; *sun* could also be taking for *roast*. But in any case where the verb retains its inherent tone, Moses translates the sentence with the meaning of the verb as observed in examples (8 c, d, e, h, i and k). These examples are repeated as (10) below with their Moses translations.

10a Mo tà púpọ̀ => I sell púpọ̀

10b Mo tà púpọ̀ ní àná => I sell in púpọ̀ yesterday

10c Mo sùn => I slept

10d Mo sùn ní ilé => I slept at home

10e Mo sùn gan-an ni => I slept was exactly

10f Ẹlẹ́mu mí tà => Ẹlẹ́mi me sell

It is observed from the examples (10) above that the senses of the verbs were appropriately translated even though there are some structural problem in the translations. For example (10c and d) are appropriately translated while others have either a word not translated like (10a, b, and f) or that the sentence structure does not represent the structure of the target sentence (10e).

4.3.2 Subject Pronoun with Progressive Aspect

According to Owolabi (2013:127), when a first or second person subject pronoun directly precede progressive aspect (*ń*), the pronoun may lose its inherent mid tone to low tone ot retain in. Moses could not recognise this difference as show in the translation of example (11).

11a Mò ń jó => Mò dancing

11b Mo ń jó => I dancing

You will observe that Moses does not recognise *Mò* in example (11a) as a Yoruba pronoun because of the change in the tone which is necessitated by the progressive aspectual marker. Example (11b) where Moses recognise *Mo* as Yoruba pronoun, the copular *am* is missing to appropriately conform to the structure of English language to give a proper translation like *I am dancing.* This means that much is still needed to be done so that for the structures of Yoruba and English could be represented by the Ngram.

4.3.3 Verb with Pronoun Object

When a verb has mid or low tone, it pronoun object must bear high tone but in a situation when the verb has high tone, the pronoun object bears mid tone (Owolabi 2013:127). This is not captured in Moses hence, the machine only recognise those pronouns with their inherent tone. In a situation when examples follows Owolabi's explanation, the computer either do not translate the pronoun or give a wrong translation. Although, there is an exception to this conclusion as regard first person singular object pronoun as observed in example 12 below:

12a Adé bú wa => Adé abuse us
b Adé lù wá => Adé to beat
c Adé gbá mi=> Adé slap me
d Adé gbà mí=> Adé save me

You will observe that the machine could recognise *mi* in spite its tonal change in (12 c and d) even though tense in the said examples are not appropriately computed. But this cannot be said about it plural counterpart *wa* as observed in example (12 a and b), the computer sees wa as separate entity because of difference in tonal change.

4.3.4 Progressive Aspect with Verb (object) and another Verb

The occurrence of a progressive aspect with two verbs co-occurring, there could be the vowel lengthening of the first verb or the object of the first verb with high bearing tone mark. This could also be retained without the vowel lengthening (see Owolabi 2013:127-128). Consider example (13) below:

13a Ó ń bú wa á lọ => he abuse us go
b Ó ń bú wa lọ=> he abuse us

You will observe that (13a) translate the notion of *lọ* which is missing in (13b) even though but translation missed the translation of progressive aspect which you give a translation as *He was abusing us while going.* This implies that the machine relies much on the available examples in the training corpus. Then, the question the readily come to mind is that, could we have corpora that will be enough for machine to learn all these tone manipulation.

4.3.5 Subject and Predicate

When the last syllable of a noun phrase bears a mid and it is directly followed by a low tone bearing constituent the mid tone of the last syllable of the Noun phrase will change to high tone. And whenever the last syllable of the Noun phrase bears low tone and it is directly followed by a mid-bearing tone predicate, the low tone of the last syllable of the Noun phrase changes to high tone as showing in example (14) below. It should also be reported that Moses does not recognise these kinds of tone manipulations.

14a Ẹlẹ́mi mí tà => Ẹlẹ́mi me sell
b Ẹlẹ́mi mi tà => my Ẹlẹ́mi sell
c Abọ́ ẹ̀wà sọnù => abọ́ ẹ̀wà lost
d Abọ́ ẹ̀wá sọnù => abọ́ ẹ̀wà lost
e òjò rọ̀ púpọ̀ => the rain rọ̀ púpọ̀
f òjó rọ̀ púpọ̀ => òjó rọ̀ púpọ̀

Example (14a) with high tone bearing syllable is translated as first person singular object pronoun while mean it has its inherent tone, it is the seen as first person possessive pronoun (my) that it is supposed to be (14b). Example (14c&d) are not recognised by Moses except the verb in the sentences because no weight is assigned to them in our training corpus. The same goes for (14 e&f) just that the computer recognise *òjò* in its inherent tone bearing form. It also has to be in lower case for it to be translated.

4.3.6 Noun and Its Qualifiers

According to (Owolabi 2013:130) When a noun has a consonant initial qualifier (nominal qualifiers and pronoun), and the last syllable of the noun is not mid tone bearing syllable, the syllable will be lengthened with mid tone as seen in example (15)

15 a Ìwée Dúdú => Ìwée Dúdú
b Ìwé dúdú => the black sexual

c ìwé dúdú => the black book

The examples like (15) show the difference between adjectival qualifiers and nominal qualifiers in Yoruba language. As observed, (15a) is not translated at all. We are still puzzled why *Ìwé* with uppercase initial is translated as sexual in seen is (15b) but when the uppercase changed to lower case, the translation was corrected as in (15c). This implies that Moses has not learn mid tone vowel lengthening like in (15a) from our training corpus. We are aware that some scholars have given explanation of the lengthened constituent different from Owolabi's position above. Our major concern here is not to support or debunk any claim but to establish that Moses as for now has not learn syllable lengthened constituents. If we consider Emphatic sentence given by Owolabi (2013:131), computer could not make a difference between the emphatic and its declarative counterpart. Consider example (16):

16 a Adé gbá mi => Adé slap me
b Adé gbá miì => Adé miì slap
c Mo ra iṣu => I bought yam
d Mo ra iṣuù =>I bought iṣuù

While (16 a and c) are declarative sentences, (16 b and d) are emphatic. As observed in the examples, the declarative sentences are translated correctly while the emphatic counterparts are not. This buttress the fact that much of tonal manipulation and vowel or syllable lengthening have not been mastered by the computer in its translations.

This proves that even though machine is modelling human language, the challenge is whether there could be enough sentences with these kinds of structure in the available literature which will give machine the necessary capacity to assign weight for such peculiar structures.

4.4 Text Editor and Tone Bearing Units

Notepad ++ serves as text editor for this study. It was observed that tone bearing units with either high or low tone especially with vowels like ọ or ẹ are separated from other constituent of words ot which they occured. This greatly affect training translation system as can be found in the *mert.out* file in the working directory. An example of such wrong translation is found below:

Translating: wọ` nyí ni ìdílé àwọn ọmọ Noa , gẹ́ gẹ́ bí ìran wọn , ní oríl ẹ̀ -èdè wọn : láti ọwọ´ àwọn wọ` nyí wá ni a ti pín oríl ẹ̀ -èdè ayé lẹ´ yìn kíkún-omi .

BEST TRANSLATION: they are family of the children of Noa|UNK|UNK|UNK , as German them , in registration : from the hands of these was then that we have shared registration cities of the world after kíkún-omi|

As observed from the extract above, *wọ̀nyí, gẹ́gẹ́, orílẹ̀*, and *lẹ́yìn* are not written together as a result of ọ and ẹ bearing high/low tone and this influences the resulting wrong translation seen in the extract. This implies that there is a need to get a text editor which would recognise Yoruba vowels with their tone bearing nature. This will help in training the translation system and thereby improve the translation outcome.

CHAPTER FIVE
SUMMARY, CONCLUSION AND RECOMMENDATION

5.0 Preamble

This chapter summaries the work so far which leads to its conclusion, thereby make some recommendation.

5.1 Summary

SMT is a widely used approach to MT. it is acclaimed that the approach performs better translation than other approaches as well as safes time and cost. The approach needs a huge bilingual corpora to learn from. It is also an approach where linguistic information is not provided before machine translates. It could then be concluded that the machine learns the linguistic information necessary for translation from the corpus huristically and mathematically. Hence, the linguistic information elicited from the corpus is could be a deviant form the common and general linguistic knowledge. The question is, how will a machine translate human natural language without the provision of any linguistic information? How will the machine learn Yoruba as a tone language? Moses, one of SMT toolkit is adopted for this work because it is an open-source tool with a mailing list where a researcher could relate with other users and experts in the field. It is also easy to adapt Moses to any language pairs. A total of 16,146 parallel sentences are used for this research. At the end of the trainings, Moses is actually translating Yoruba to English though with some considerable errors. This brings to fore two conclusions that: linguistics information like part of speech tagging are to be provided in the training corpus to reduce the errors. It is also needful to provide more training data as the system demands when compared with available traning data for language pairs like French-English, German-English and so on.

It was found that Moses learns Yoruba words with their inherent tones but have not actually learn tone variations and manipulations as its related to Yoruba language sentence structures.

5.2 Conclusions

Based on the reseach questions, experiement, presentation and analysis, below are some of our conclusions:

That mahine cannot acquire human language but can model it since language acquisition capacity is restricted to humans alone. It is needful however, to say clearly that human natural language could be modelled by a machine if necessary linguistic information like part of speech tagging, semantic and pragmatic annotation are provided in the training corpus.

That machine irrespective of what it is used for, remains a tool for human assistance hence its performance depend absolutely on human manipulation. Therefore, machine can translate but not as human. Although is it plausible to think that machine can translate like humans if and only if all possible and imaginable sentences can be available for machine to learn from. To us, all possible human sentences cannot be reduce to text because language has much to do with experience, innovation, reasoning and exposure. Furthermore, it may be difficult to reduce language sentence structures to mathematical rules.

5.3 Recommendations

It is not enough to say that MT is not translating African languages appropriately thereby limiting to human translation. It is recommended that efforts should be directed towards translating most of the available literary text if not all so as to have enough training corpus in order to expand the frontiers of the activities of MT in relation to African languages to at least produce a first draft for its users before appropriate necessary editing. In other words, literary texts may have their unique challenges, in our opinion, they could be used as a means to build corpora for this kind of exercise. Hence efforts should be channeled towards corpora resource building. When we have enough corpora, with improved technical know-how, the machine should be able to give close to reasonable translation. This will save time and cost. For example it took computer an average of 0.7 seconds to translate a sentence while it took humans an average of 4 to 5 minutes to do same. Human translator will charge a reasonable amount for the service of translation while Google and Ibadan SMT charged nothing. It would also be an avenue to contribute to the ongoing discussion over the global space on the internet; that Africans with their language can showcase their potentials without any language limitations whatsoever. Scholars should, therefore, come together with their expertise to achieve the necessary needed improvement in terms of resource building and technicalities.

African governments need to ensure that projects in relation to technology that are language focused should be sponsored no matter the small financial resources in the region. Africa, as a continent stand to gain much if her languages are promoted via technology. More importantly, research of this nature will give Africa more visibility globaly when an ordinary African without English language competence can assess internt effortlessly nd thereby improve her economy.

REFERENCES

Agarwal, A. and Lavie A. 2008. METEOR, M-BLEU and M-TER: Evaluation Matrics for High-Correlation with Human Ranking of Machine Translation Output. *Proceeding of the Third Workshop on Statistical Machine Translation*. pp 115-118 Columbus, Ohio USA @ June, 2008 Association of Computational Linguistics.

Ajadi. G.A. 2005. *The Forest Of God: Annotated Translation of D.O. Fagunwa's Igbó Olódùmarè*. 3rd ed. Ilorin. Bamitex Printing and Publishing Co.

Akinwale O. I., Adetunmbi A. O., Obe O. O., Adesuyi A. T.. 2015. Web-Based English to Yoruba Machine Translation. *International Journal of Language and Linguistics*. Vol. 3, No. 3, pp. 154-159. doi: 10.11648/j.ijll.20150303.17

Awobuluyi O. 2010. The Role of Linguistics in Nation Building: Lecture in Honour of Ayo Bamgbose, Delivered at the University of Ibadan

Banerjee, S. and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgment. *Proceeding of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measure for Machine Translation and/or Summarization*. Ann Arbor MI 65-72

Bar-Hillel Y. 1954. Some Linguistic Problems Connected with Machine Translation. *Philosophy of Science*, vol.20, pp 217-225

Chan H.Y and Rosenfeld R. 2012. Discriminative Pronunciation Learning for Speech Recognition for Research Scarce Languages. *ACM 978-1-4503-1262-2/12/03*

Chéragui, M.A. 2012. Theoretical Overview of Machine Translation; *Proceeding ICWIT*

Chomsky N, 2006. *Language and Mind*. Cambridge. Cambridge University Press

De Pauw G., Wagacha P. and Schryver G. 2011. Exploring the SAWA Corpus: Collection and Deployment of a Parallel Corpus English and Swahili. Language Resources and Evaluation. Vol. 45, No. 3, (Summer) pp 331-344. http://www.jstor.org/stable/41486046 Assess on 02.04.2012

Egbokhare F. 2011. The Sound of Meaning. An Inaugural Lecture Delivered at the University of Ibadan on July 14,

Fagunwa D.O. 2005a. *Ògbójú Ọdẹ Nínú Igbó Irúnmọlẹ̀*. 20th ed. Ibadan. Nelson Publishers Limited

__________. 2005b. *Igbó Olódùmarè*. 20th ed. Ibadan. Nelson Publishers Limited

__________. 2005c. *Ìrìnkèrindò Nínú Igbó Elégbèje*. 11th ed. Ibadan. Nelson Publishers Limited

__________. 2005d. *Àdììtú Olódùmarè*. 12th ed. Ibadan. Nelson Publishers Limited

Fernandes L. 2008. Corpora in Translation Studies: Revisiting Baker's Typology. https://periodicos.ufsc.br/.../7690&sa=U&ved=0CB8QFjACahUKEwjtprCShODHAhX...Retrieved on 1/8/2016

Gavrila M. and Vertan C. 2011. Training Data in Statistical Machine Translation – The More, the Better? *Proceedings of Recent Advances in Natural Language Processing*, pp 551-665, Hissar, Bulgaria, 12-14 September

Hassan H. 2009. A Lexical Syntax for Statistical Machine Translation. A PhD Thesis submitted to the School of Computing, Faculty of Engineering and Computing; Dublin City University

Hutchins, J. 2001. Machine Translation and Human Translation: in Competition or in Complementation? *International Journal of Translation* 13, 1-2 Jan-Dec 2001, pp.5-20.

___________ 2007. Machine Translation: A Concise History. In *Computer Aided Translation: Theory and Practice*, C. S. Wai (ed.). Chinese University of Hong

__________ and Somers H.L 2009. *An Introduction to Machine Translation*. London: Academic Press Limited

Isola A. 2001. *Aké Ní Ìgbà Èwe*. Ibadan: BookCraft LTD

Kamssu, J.; Siekpe, J.S. and Elizy, J.A. 2004 Shortcomings to Globalization: Using Internet Technology and Electronic Commerce in Developing Countries. *The Journal of Developing Areas*, Vol. 38, No. 1 pp. 151-169. Retrieved Jan 25, 2012 from http://www.jstor.org/stable/20066700

Koehn, P. 2004. Statistical Significance Test for Machine Translation Evaluation. Proceeding of EMNLP 2004. Lin D and Wu D Eds 1-8

_______ 2010. *Statistical Machine Translation*, Cambridge. Cambridge University Press

_______ 2012. Statistical Machine Translation System: User Manual and Code Guide. Retrieved Mar 12, 2012 from www.moses.org.

Lavie, A. 2014. Automated Metrics for MT Evaluation. Pp 11-731; Feb 20, 2014

Lita, L.V., Ittycheriah, A., Roukos, S. and Kambhatla, N. 2003. tRuEcasIng. In Hinrich, E. and Roth, D. editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* pp152-159

Lopez A. 2008. Statistical Machine Translation. *ACM Computing Surveys*. Vol 40, No 3, Article 8, August 2008.

Mearne, M. and Way, A. 2011. Statistical Machine Translation: A Guide for Linguists and Translators. *Language and Linguistics Compass* pp 1-21

Obafemi O. 2009. *The Mysteries of God: A Translation of D.O. Fagunwa's Adiitu Olodumare*. Ibadan. Nelson Publisher Limited.

Odoje C.O 2010. The Role of Syntax in the Yoruba-English Rule-Based Machine Translation. A Master Project Submitted to the Department of Linguistics and African Languages, University of Ibadan

________. 2013a. The Peculiar Challenges of Statistical Machine Translation to African Languages. *Issues in Contemporary African Linguistics: Festschrift No 11*. Ndimele, O.M., Yuka, L.C, and Ilori, J.F Eds. The Linguistics Association of Nigeria in Collaboration with N and J Grand Orbit Communication Ltd. Port Harcourt.

_________. 2013b. Language: A Catalyst for National Development. *Linguistics and Glocalisation of African Languages for Sustainable Development: A Festschrift in Honour of Prof. Kola Owolabi*. Adegbite, W., Ogunsiji, A., and Taiwo O. Eds. Universal Akada Books Nigeria Limited. Chapter 2: 34-47

________. and Akinola S 2013. Exploring the Challenges of SMT Projects for Resource Scarce Languages in Africa: A case Study of English-Yoruba Machine Translation. *RALL: Research in African Languages and Linguistics* Vol 12, pp 84-94

________. 2014. Investigating Language in the Machine Translation: Exploring Yoruba-English Machine Translation as a Case Study. SamaraAltinguo e-journal Issue 05, pp4-11. Retrieved Feb 13, 2015 from www.samaraaltinguo.com

Omary Z. and Mtenzi F. 2010. Machine Learning Approach to Identifying the Dataset Threshold for the Performance Estimators in Supervised Learning. *International Journal for Infonomic* (IJI) Vol 3, Issue 3, September pp 314-325

Osundare N. 1995. Caliban's Gamble: The Stylistic Repercussions of Writing African Literature in English in *Languge in Nigeria: Essays in Honour of Ayo Bamgbose*

Owolabi k. 2013. *Ìjìnlẹ̀ Ìtúpalẹ̀ Èdè Yorùbá: Fònẹ́tîìkì àti Fonọ́lọ́jì*, Ibadan. Universal Akada Books Nigeria Limited.

Papineni, K., Roukos, S., Ward, T., and Zhu W. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceeding of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL) Philadelphia, July 2002 pp 311-318.

Schleicher A.T.F. 1998. *Jẹ́ k'á ka Yorùbá*. New Haven; Yale University Press

Shen G. 2010. Corpus-Based Approach to Translation Studies. *Cross-Cultural Communication*. Vol 6 No 4 pp 181-187

Slocum, J. 1985. A Survey of Machine Translation: Its History, Current Status and Future Prospects. *Computational Linguistics*, Vol 11 Number 1, pp. 1-17.

Snover, M., Dorr, B., Schwaetz, R., Micciulla, L., and Makhoul, J. 2006. A study of Translation Edit Rate with Targeted Human Annotation. *Proceeding of the Conference of the Association for Machine Translation in America* (AMTA) 223-231

Soyinka W. 1982. *Forest of A Thousand Daemons*. New York. Random House

________. 2000. Ake: *The Years of Childhood*. 3rd. London. Methun Publishing Limited

________. 2010. In the Forest of Olodumare: A Translation of D.O. Fagunwa's Igbo Olodumare. Ibadan. Belson Publishers Limited

Stevick E.W and Aremu O. 1963. *Yoruba Basic Course*. Washington DC. Foreign Service Institute, Department of State www.asiaonline.com

Yusuf O. 2011. *Basic Linguistics for Nigerian Languages* 3rd ed. Ijebu-Ode. Shebiotimo Publications

Appendix 1

```
# testrun moses manual for Yoruba->English

# Tokenising

# English

~/moses/scripts/tokenizer/tokenizer.perl -l en \  < ~/corpus/training/English.yor-en.en
\  > ~/corpus/yoruba_english.yor-en.tok.en

  # Yoruba

 ~/moses/scripts/tokenizer/tokenizer.perl -l yr \  < ~/corpus/training/Yoruba.en-
yor.yor \  > ~/corpus/yoruba_english.yor-en.tok.yor

  # Truecase-training

# English

  ~/moses/scripts/recaser/train-truecaser.perl \--model ~/corpus/truecase-model.en --
corpus \~/corpus/yoruba_english.yor-en.tok.en

# Yoruba

~/moses/scripts/recaser/train-truecaser.perl \ --model ~/corpus/truecase-model.yor --
corpus \ ~/corpus/yoruba_english.yor-en.tok.yor

# Truecasing

# English

~/moses/scripts/recaser/truecase.perl \ --model ~/corpus/truecase-model.en \
< ~/corpus/yoruba_english.yor-en.tok.en \ > ~/corpus/yoruba_english.yor-en.true.en

# Yoruba

~/moses/scripts/recaser/truecase.perl \ --model ~/corpus/truecase-model.yor \
< ~/corpus/yoruba_english.yor-en.tok.yor \ > ~/corpus/yoruba_english.yor-en.true.yor

# Cleaning limit 80 words

~/moses/scripts/training/clean-corpus-n.perl \
~/corpus/yoruba_english.yor-en.true yor en \
~/corpus/yoruba_english.yor-en.clean 1 80


# Language Model training

mkdir ~/lm

cd lm !!!
```

```
~/moses/bin/lmplz -o 3 <~/corpus/yoruba_english.yor-en.true.en >
yoruba_english.yor-en.arpa.en

# Then we should Binarize

~/moses/bin/build_binary \ yoruba_english.yor-en.arpa.en \
yoruba_english.yor-en.blm.en

# Check the Language model by querying it
$ echo "Èmi fẹ́ràn owó" \
> | ~/moses/bin/query yoruba_english.arpa.en.blm.en

Èmi=2253 1 -4.89647     fẹ́ràn=141 1 -3.32892    owó=258 1 -3.39753      </s>=2 1 -
2.33102       Total: -13.9539 OOV: 0
Perplexity including OOVs:      3079.53
Perplexity excluding OOVs:      3079.53
OOVs:   0
Tokens: 4
Name:query      VmRSS:4256 kB   RSSMax:4284 kB  user:0.015      sys:0.015
CPU:0.03       real:0.0242441

# Training the Translation System

mkdir ~/working
cd ~/working
nohup nice ~/moses/scripts/training/train-model.perl -root-dir train \
-corpus ~/corpus/yoruba_english.yor-en.clean \
-f yor -e en -alignment grow-diag-final-and -reordering msd-bidirectional-fe \
-lm 0:3:$HOME/lm/yoruba_english.yor-en.blm.en:8 \
-external-bin-dir ~/moses/tools >& training.out &
```

This may take some time

```
##########################################################
Tuning
cd ~/corpus

~/moses/scripts/tokenizer/tokenizer.perl -l en \
  < ~/corpus/training/EnglishYorubaTuning.yor-en.en \
  > ~/corpus/EnglishYorubaTuning.yor-en.tok.en


# Yoruba

 ~/moses/scripts/tokenizer/tokenizer.perl -l en \
  < ~/corpus/training/EnglishYorubaTuning.yor-en.yor \
  > ~/corpus/EnglishYorubaTuning.yor-en.tok.yor


 # Truecasing of Tuning data
# English

~/moses/scripts/recaser/truecase.perl \
--model ~/corpus/truecase-model.en \
< ~/corpus/EnglishYorubaTuning.yor-en.tok.en \
> ~/corpus/EnglishYorubaTuning.yor-en.true.en


# Yoruba

~/moses/scripts/recaser/truecase.perl \
--model ~/corpus/truecase-model.yor \
< ~/corpus/EnglishYorubaTuning.yor-en.tok.yor \
> ~/corpus/EnglishYorubaTuning.yor-en.true.yor


# Now go back to the directory we used for training, and launch the tuning process:

# cd ~/working

nohup nice ~/moses/scripts/training/mert-moses.pl \
~/corpus/EnglishYorubaTuning.yor-en.true.yor  ~/corpus/EnglishYorubaTuning.yor-en.true.en \
~/moses/bin/moses train/model/moses.ini --mertdir ~/moses/bin/ \
&> mert.out & --decoder-flags="-threads 4"


# Testing

~/moses/bin/moses -f ~/working/mert-work/moses.ini

# Binarizing (takes time)

 mkdir ~/working/binarised-model

cd ~/working



 ~/moses/bin/processPhraseTableMin \
-in train/model/phrase-table.gz -nscores 4 \
-out binarised-model/phrase-table


~/moses/bin/processLexicalTableMin \
-in train/model/reordering-table.wbe-msd-bidirectional-fe.gz \
-out binarised-model/reordering-table

#########################################################
Now we need another corpus different from what we have used so far
#########################################################
#cd corpus
```

```
~/moses/scripts/tokenizer/tokenizer.perl -l en \
< EnglishYorubaBTest.yor-en.en > EnglishYorubaBTest.yor-en.tok.en

~/moses/scripts/tokenizer/tokenizer.perl -l yor \
< EnglishYorubaBTest.yor-en.yor > EnglishYorubaBTest.yor-en.tok.yor

~/moses/scripts/recaser/truecase.perl --model truecase-model.en \
< EnglishYorubaBTest.yor-en.tok.en > EnglishYorubaBTest.yor-en.true.en

~/moses/scripts/recaser/truecase.perl --model truecase-model.yor \
< EnglishYorubaBTest.yor-en.yor > EnglishYorubaBTest.yor-en.true.yor

#cd working
~/moses/scripts/training/filter-model-given-input.pl \
filtered-EnglishYorubaBTest.yor-en mert-work/moses.ini
~/corpus/EnglishYorubaBTest.true.yor \
-Binarizer ~/moses/bin/processPhraseTableMin


nohup nice ~/moses/bin/moses \
-f ~/working/filtered-EnglishYorubaBTest.yor-en/moses.ini \
< ~/corpus/EnglishYorubaBTest.yor-en.true.yor \
> ~/working/EnglishYorubaBTest.yor-en.translated.en \
2> ~/working/EnglishYorubaBTest.yor-en.out
~/moses/scripts/generic/multi-bleu.perl \
-lc ~/corpus/EnglishYorubaBTest.yor-en.true.en \
< ~/working/EnglishYorubaBTest.yor-en.translated.en
```

Appendix 2

DEPARTMENT OF LINGUISTICS AND AFRICAN LANGUAGES
UNIVERSITY OF IBADAN

Questionnaire

What is this study about?
This is part research project being conducted by ODOJE Clement Oyeleke, a PhD student of the Department of Linguistics and African Languages, University of Ibadan. We are inviting you to participate in this research project as one of our potential human translation evaluator. Your valued opinion and support will assist us in evaluating Google translate (machine translation) in comparison with Human translation so as to find linguistic explanation to both bad and good translations.

What will I be asked to do if I agree to participate?
If you agree to participate in this research, you will be asked to evaluate translations of some Yorùbá sentences into English of both Machine and Humans. Your objectivity is highly necessary here. The source sentence is Yorùbá while Target sentence is English. You are to evaluate accuracy which is the appropriateness of translation and fluency which is how good the sentence is in the target language. Your ratting is 1-5. 5 = excellent; 4 = very good; 3 = good; 2 = poor; 1 = very poor. You are to write the appropriate ratting based on your judgment is the provided space as exemplified in the table below:

Yorùbá sentence	Machine Translation	Accuracy of translation	Fluency of translation in the target language	Human Translation	Accuracy of translation	Fluency of translation in the target language
Olú rí mi	My capital	1	1	Olu sees me	4	3

Personal Information

1. Age in Years _______________
2. Education: Secondary Education [ ], College of Education [ ],
   Undergraduate [ ], Postgraduate [ ], others [ ].
3. Marital Status: Single [ ], Married [ ], Widowed [ ],
   Separate /Divorced [ ]
4. Religion: Christianity [ ], Islam [ ], Traditional [ ] others [ ]

Questionnaire

S/N	Yoruba Sentences	English Sentence Translated by Machine			English Sentence Translated by Human		
			Adequacy of translation	Fluency of translation in the target language		Adequacy of translation	Fluency of translation in the target language
1	Olú rí mi	My capital			Olu sees me		
2	Olú rí ẹ/ọ	Capital			Olu sees you		
3	Olú rí i	Capital found			Olu sees him/her		
4	Olú rí wa	Our Capital			Olu sees us		
5	Olú rí yín	Capital			Olu sees you		
6	Olú rí wọn	Their capital			Olu sees them		
7	Mo rí Olú	I see the Divine			I saw Olú		
8	O rí Olú	You see Emperor			You saw Olu		
9	Ó rí Olú	He saw the Divine			He saw Olú		
10	A rí Olú	A Capital			We saw Olú		
11	Ẹ rí Olú	Capital			You saw Olú		
12	Ẹ rí Olú	Their Capital			They saw olú		
13	Èmi rí Olú	I see the Divine			I saw Olu		
14	Ìwọ rí Olú	You see the Divine			You saw Olú		
15	Òun náà rí Olú	He saw the Divine			He too saw Olú		
16	Àwa pàápàá rí Olú	We even saw the Emperor			We too/even saw Olú		
17	Ẹ̀yin gan rí olú	Emperor eggs			You in particular saw Olú		
18	Ẹ̀yin rí Olú	Emperor eggs			You saw Olú		
19	Àwọn náà rí Olú	The capital			They too saw Olú		
20	Olè jí owó rẹ	Steal your money			A thief stole your money		
21	Olè jí owó rẹ̀	Steal your money			A thief stole his money		
22	Olè jí owó wa	Steal our money			A thief stole our money		
23	Olè jí owó yín	Steal your money			A thief stole your money		
24	Olè jí owó wọn	Steal their money			A thief stole their money		

25	Tiwa dà?	Our?			Where is ours		
26	Tèmi dà?	I like?			Where is mine		
27	Tìrẹ dà?	Yours?			Where is yours		
28	Tiyín dà?	Yours?			Where is yours		
29	Tiwọn d à?	Them?			Where is theirs		
30	Èyí dára	This is			This is good/this one is good		
31	Ìyẹn dára	The best			That is good / That one is good		
32	Ìwọnyí dára	These are good			These ones are good / These are good		
33	Ìwọnyẹn dára	These are good			Those are good/ Those ones are good		
34	Olú wà ní ibí	Capital in place			Olu is here		
35	Olú wà ní ibẹ	Capital in place			Olu was there		
36	Táyé wà ní ọhún	Entertaining us			Táyé was in the place		
37	Táyé dé ní àná	Road yesterday			Táyé arrived yesterday		
38	Olú kò dé ní àná	Capital and yesterday		-	Olú did not arrive yesterday		
39	Olú ti dé láti àná	Capital has come from yesterday			Olu had arrived yesterday		
40	Olú máa dé ní ọla	Capital			Olu will arrive tomorrow		
41	Olú kò níí dé ní ọla	Capital do not come at you			Olu will not arrive tomorrow		
42	Olú kò tîì dé	Capital had not yet come			Olu had not arrived		
43	Olú n bọ	Capital coming			Olu is coming		
44	Ó ti n bọ báyìí-báyìí	He is coming now-now			Olu is coming right now		
45	Mo gbọ pé Olú dé	I heard the emperor arrived yesterday			I heard that Olu arrived yesterday.		

46	Ó yẹ kí Olú dé ní ọ̀la	it should come in tomorrow emperor			Olu is supposed to arrive tomorrow		
47	Mo fẹ kí ó dé ní ọ̀la	I want you to come in tomorrow			I want Olu to arrive tomorrow		
48	Mi ò mọ bóyá Olú dé ní ọ̀la	I do not know whether to come in tomorrow emperor			I doubt if Olu will arrive tomorrow		
49	Ó dára pé Olú wá	It is good that the emperor			It is good that Olu comes		
50	Ó dára kí Olú wá	Well capital			It is good that olu comes		
51	Kò dára rárá kí ó	Not good at all for you to come			It is not good at all that he would not come		
52	Mo ní kí Olú lọ, kí sì tètè padà	Emperor i have to go, and quickly returned			I commanded Olu to go and come back immediately		
53	Olú ní kí n jáde	I capital			Olu said I should go out		
54	Olú ní kí o jáde	Capital			Olu said you should go out		
55	Olú ní kí ó jáde	Capital			Olu said he/she should go out		
56	Táyé ní kí ó jáde	Scene in spring			Taye said He/she should go out		
57	Olú ní kí ẹ jáde	Capital out			Olu said that you should go out		
58	Olú ní kí wọn jáde	Their capital in order			Olu said they should go out		

59	Nj́ẹ́ Ìbíkúnlé dé ní	Then the election came in yesterday?			Did Ibikunle arrive yesterday?		
60	Ṣé Ṣẹ́gun dé ní ànà́?	Victory came in yesterday			Did Segun arrive yesterday?		
61	Àbí Olú dé ní àná?	Emperor or yesterday			Is it true that Olu arrive yesterday ?		
62	Olú dé ní àná bí?	Capital yesterday			Did Olu arrive yesterday?		
63	Olú dà?	Capital?			Where is Olu?		
64	Olú nkọ?	Capital?			What about Olu/ Where is Olu?		
65	Ibo ni Olú wà?	Where is emperor?			Where is Olu?		
66	Ọjọ wo ni Olú lọ?	What date is the emperor?			Which day did Olu left?		
67	Ìgbà wo ni ó máa padà?	What time is it?			When is he returning?		
68	Bá wo ni Olú se máa padà?	What if the emperor’s return?			How is Olu returning?		
69	Ó máa gun kẹkẹ ni, àbí ó máa wọ mọtò?	He will ride in, it will go into the car?			Will you ride a bike or drive a car?		
70	Ibo ni Olú ti nbọ?	Where capital is coming?			Where is Olu coming from?		
71	Kí ni Olú lọ se?	What is capital?			What has Olu gone to do?		
72	Ta ni Olú lọ kí?	Who is the emperor?			How did Olu go out to greet		
73	Kí ni o torí sọ bẹẹ?	What do you say?					
74	Kí ni o se sọ bẹẹ?	What do you say?			Why did you say so?		
75	Kúrò ní ọdọ mi!	Away from me!			Leave me!		
76	Ẹ kúrò ní ọdọ mi!	Get away from me!			You all should leave me!		
77	Ìwọ, jáde wá sí ibí yìí!	You, come here!			You, come out here!		

78	Èyin, ẹ jáde wá sí ibí yìí!	You, come here!			You, come out here!		
79	Èyin ọmọ wọnyí, mo ní kí ẹ jáde wá sí ibí!	Dear children, I have to get out of here!			You Children, I said you should come out here		
80	Má na ọmọ yẹn mọ́!	Not that the children know!			Stop beating the kid		
81	Àní ìyẹn tó, ẹ má nà án mọ́	Infact, do not let know			I said that is alright, stop beating him/her		
82	Èmi ni Olú ń kí	I am emperor"s			Olu is greeting me		
83	Ìwọ ni Olú n kí	O divine			Olu is greeting you/ it is you Olu is greeting		
84	Èyin ni Lékè ń kí	Above			It is you Leke is greeting/ Leke is greeting you		
85	Òun ni Lékè ń kí	Above			His is the one Leke is greeting/ Leke is greeting him		
86	Àwa ni Lékè ń kí	Above			Leke is greeting us/ It is we that Leke is greeting		
87	Àwọn ni Lékè ń kí	Above			Leke is greeting them, it is they that leke is greeting		
88	Dàda ni Lékè ń kí	Workout above			It is Dada that Leke is greeting/ Leke is greeting Dada		
89	Dàda ni ó ń kí Olú	It is well to capital			It is Dada that is greeting Olu		
90	Dàda ni ó ń kí Ṣẹ́gun	It is well to victory			It is Dada that is greeting Tolu		
91	Èmi ni mo ń kí Tolú	I Perseverance			It is I that is greeting Tolu		
92	Ìwọ ni ó ń kí Tolú.	You have to perseverance			It is you that is greeting Tolu		

93	Òun ni ó ń kí Tope	He is dead			He is the one greeting Tope		
94	Àwa ni a ń kí Olú	We have capital			We are the one greeting Olu		
95	Èyin ni ẹ ń kí Olú	You are the divine			You are the one greeting Olu		
96	Àwọn ni wọn ń kí Olú	The capital			They are the one greeting Olu		
97	Dàda ni ó jí mọtò Olú	Workout stolen car capital			Dada is the one who steals Olu's car		
98	Mọtò Olú ni Dàda jí	Emperor car is properly restored			It is Olús car that Dada stole		
99	Olú ni Dàda jí mọtò rẹ	Your Car is properly raised capital			It is Olu who Dada stole his car		
100	Èmi ni olè jí mọtò rẹ	I steal your car			It is I whose car was stolen by the thief		
101	Ìwọ ni olè jí mọtò rẹ	You steal your car			You are the one whose car was stolen		
102	Òun ni olè jí mọtò rẹ	He is a thief stole your car			He is the one whose car was stolen		
103	Àwa ni olè jí mọtò rẹ	We steal your car			We are the one whose car was stolen		
104	Èyin ni olè jí mọtò rẹ	You a thief stole our car			You are the one thief stole their car		
105	Ibo ni olè ti jí mọtò Olú?	Where the car had be stolen emperor			Where did the thief stole Olu's car?		
106	Ọjà ni olè ti jí mọtò Olú	Market has be stolen car capital			It is in the market that the thief stole Olu's car		
107	Sé ó dájú pé jíjí ni olè jí mọtò rẹ?	Are you sure that raising the thief steal your car?			Is it true that your car was actually stolen?		
108	Àbí olè kàn yá a lò láásán ni?	Or a thief use láásán?			Or that a thief just borrows it.		

109	Dájú-dájú, jíjí ni ó jí i, níwọn ìgbà tí kò ti tọrọ rẹ lò	Sure-sure, cheating is raised, since no request your			Surely, the car was stolen for the fact that he did not request for it.		
110	Mọtò tí Olú rà dà?	Moto capital bought			Where is the car Olu bought		
111	Olú tí ó ra mọtò dà?	Capital to buy insurance?			Where is the Olu who bought the car		
112	Kí ni wọn ń pe ibi tí Olú ti ra mọtò rẹ?	What is called the Capital have bought your car?			What is the place Olu bought his car from called		
113	Èmi tí ó ń sọrọ yìí ti ra mọtò	I was talking to this purchase insurance			I that is speaking have bought a car		
114	Èyin tí ẹ fẹ ra mọtò, ẹ na ọwọ sókè!	If you want to buy a car, your hands up!			If you want to buy a car, raise up your hands		
115	Kò yẹ kí o na ọmọ náà tó bẹẹ; nínà tí o nà án ti pọ jù lójú tèmi.	If you want to buy a car your hands up!			You don’t need to raise your hand to that extent, that at to me is too much		
116	Mo dé ní ìgbà tí wọn ń jẹun	I came in when they eat			I arrive when they were eating		
117	Gbàrà tí eré bẹrẹ ni òjò ńlá kan dé	Once the games start on a greats			Once the game started, there came a heavy rain		
118	Bí wọn se bẹrẹ eré ni òjò dé	They start the day			Once the game started, the rain began		
119	Èyìn tí òjò dá ni eré tún bẹrẹ	If the day start			After the rain stopped, the game began		
120	Kí òjò tó bẹrẹ ni mo ti wọlé ní tèmi	Before the rain started, I have come in part			I entered before the rain started		
121	Bí òjò bá tètè rọ, ìyàn kò níí mú.	A rain delay, and do not bring famine			If rain starts early, there won’t be famine.		
122	Bí òjò kò bá tètè rọ, ìyàn máa mú gan-an	If rain does not easily persuaded, famines			If rain begins early, there will be famine		

123	Bí òjò rọ, déédé; bí kò sì rọ, déédé náà ni	If rain, completely; And if not encouraged, completely			It is fine whether it rain or not		
124	Òjò ì báà tètè rọ, ìyàn máa mú sẹ!	Rain could be easily encouraged, famines!			Even though it rain early, I said there will be famine		
125	Òjò ì báà máà tètè rọ, ìyàn kò níí mú sẹ!	Rain could be easily encouraged. To make famine			Even though rain may not fall, there won’t be famine		
126	Ò báà sunkún jù bẹẹ, mi ò níí fún ẹ.	You cry more. I do not mind for you			Even if you cry, I will not give you		
127	Ẹ jẹ kí a sisẹ, kí a ba lè ní owó lọwọ!	Make a tape, so that we can have money!			Let’s work so that we could have money		
128	Ẹ jẹ kí a sisẹ, kí ìyà má ba à jẹ wá!	Let us work will not suffer if we are looking for!			Let’s work to avoid poverty		
129	Bí ó tilẹ jẹ pé òjò tètè rọ, ìyàn pàpà mú!	Even if it is vulnerable to rain, and forced famine!			Even though it rained early, there was still famine		
130	Bí ó tilẹ jẹ pé òjò kò tètè rọ, ìyàn kò mú!	although that is not easily persuaded, come!			Even though it did not rain early, there was no famine		
131	Mo mọ ilé Olú	I know capital			I know Olu’s house		
132	Mo mọ ìyàwó Dàda	I knew her well			I know Dada’s wife		
133	Ìbàdàn ni ìlú Adélabú	Ìbàdàn is big cities			Ibadan is Adelabu’s town		
134	Mo wọ ẹwù funfun	I dress in white			I wore a white cloth		
135	Mo wọ ẹwù pupa	I wear red garment			I wore a red cloth /garment		
136	Mo wọ asọ tuntun	I wear new			I wore a new clothe/garment		
137	Mo sá asọ tútù	I run her wet			I spread a wet clothe		
138	Mi ò sá asọ gbígbẹ	I do not run dry cloth			I did not spread a dried clothe		
139	Mo mu ẹkọ gbígbóná	I play intense training			I drank a hot pap		

140	Mo ka ẹni kan; ẹni méjì; ẹni mẹta; ẹni mẹrin, ẹni márùnún; ẹni mẹfà; ẹni méje; ẹni mẹjọ; ẹni mẹsànán; ẹni mẹwàá; ogún ẹni; ọgbọn ẹni; ọgọrùnún ẹni.	I read a person; two; parties; four, starving; six; one seven; one digit; adequate; one half; twenty one; thirty one; ọgọrùnún.			I counted a person; two people; three people; four people; five people; six people; seven people; nine people; ten people; twenty people; thirty people; hundred people		
141	Wọn pe ẹni kínní; ẹni kejì; ẹni kẹta; ẹni kẹrin; ẹni karùnún; ẹni kẹfà; ẹni keje; ẹni kẹjọ; ẹni kẹsànán; ẹni kẹwàá.	They called the first one; the second one; third; the fourth one; karùnún; sixth; seventh; one eighth; kẹsànán; one tenth.			They called first person; second person; third person; fourth person; fifth person; sixth person; seventh person; eighth person; ninth person, tenth person.		
142	Olú lọ kí Òjó	Capital			Olu went to greet Òjó		
143	Ó lọ bèrè àláfíà rẹ.	He began peace			He went to ask of his welfare		
144	Wọn máa wá kí wa.	They would come to us			They will come to greet us		
145	Wọn máa wá bèrè àláfíà wa	They will begin peace			They will come too ask of our welfare		
146	Mo he owó ní ọjà.	I have that money in the market			I picked up money in the market		
147	Wọn fẹ wá ní ọla	They want to come in tomorrow			They want to come tomorrow		
148	Mo rí owó náà ní ilẹ	I found the money in the			I saw the money on the ground		
149	Ilẹ ni mo ti rí i	I have seen			I saw it on the ground		
150	Mo sọ owó náà sí àpò	I made money in the pocket			I kept the money in the pocket		
151	Àpò ni mo sọ ọ sí	I told you to bag			I kept it in the pocket		

152	Kí ni o ti se owó náà?	What money do you have?			How did you spend the money		
153	O ná an, àbí o fi pamọ?	You spend it, or do you save?			Did you spend the money or keep it		
154	Kò yé mi bí mo se se é mọ	I did not understand how I did it			I don’t remember how I did it again		
155	Bóyá ó bọ sọnù ní àpò mi ni	Perhaps it was lost in my bag			Perhaps it dropped from my pocket		
156	Olú kò dé ibẹ rí	Capital is not there			Olu has never been there		
157	Olú kò dé ibẹ mọ	Capital and there			Olu		
158	Olú kò wá rárá	Capital did not come at all.			Olú did not come at all		
159	Olú kò wá sá	Capital does not run			Olu did not come		
160	Olú kò wá sẹ!	Our capital!			Olu did come I said		

Appendix 3

ASIEN-AFRIKA-INSTITUTE OF THE UNIVERSITY OF HAMBURG

QUESTIONNAIRE

What is this study about?

This is part of an on-going research studying human and machine translations. We are inviting you to participate in this research project as one of our potential human translation evaluator. Your valued opinion and support will assist us in evaluating Google Translate (machine translation) in comparison with human translation so as to find linguistic explanation to both bad and good translations as well as comparing the output with other languages.

What will I be asked to do if I agree to participate?

If you agree to participate in this research, you will be asked to evaluate translations of English sentences that were translated into German. Your objectivity is highly necessary here. You are to evaluate accuracy which is the appropriateness of translation and fluency which is how good the sentence is in the target language. Your rating is 1-5. 5 = excellent; 4 = very good; 3 = good; 2 = poor; 1 = very poor. You are to write the appropriate rating based on your judgment as exemplified in the table below:

English	Machine Translation	Adequacy	Fluency	Human Translation	Adequacy	Fluency
My name is Olu	Mein Name ist Olu	3	3	Mein Name ist Olu	4	3

Personal Information

1. Age in Years _______________
2. Education: Secondary Education [ ], College of Education [ ], Undergraduate [ ], Postgraduate [ ], others [ ].
3. Marital Status: Single [ ], Married [ ], Widowed [ ], Separate/Divorced [ ]
4. Religion: Christianity [ ], Islam [ ], Traditional [ ] others [ ]

English	**Google translate German**	**Adequacy**	**Fluency**	**Human Translation**	**Adequacy**	**Fluency**
Olu sees me	Olu sieht mich			Olu sieht mich		
Olu sees you	Olu sieht dich			Olu sieht dich/Ihnen		
Olu sees him/her	Olu sieht ihn/sie			Olu sieht ihn/sie		
Olu sees us	Olu sieht uns			Olu sieht uns		
Olu sees them	Olu sieht sie			Olu sieht sie		
I saw Olú	Ich sah Olu			Ich sah Olu		
You saw Olu	Sie haben gesehen, Olu			Du sahst Olu/Sie sahen Olu		
He saw Olú	Er sah, Olu			Er sah Olu		
We saw Olú	Wir sahen Olu			Wir sahen Olu		
They saw olú	Sie sahen Olu			Sie sahen Olu		
He too saw Olú	Auch er sah Olu			Auch er sah Olu		
You in particular saw Olú	Sie insbesondere Säge Olu			Insbesondere sie sahen/Du sahst Olu		
A thief stole your money	Ein Dieb stahl Ihr Geld			Ein Dieb stahl ihr/dein Geld		
A thief stole his money	Ein Dieb stahl sein Geld			Ein Dieb stahl sein Geld		
A thief stole our money	Ein Dieb stahl unser Geld			Ein Dieb stahl unser Geld		
A thief stole your money	Ein Dieb stahl Ihr Geld			Ein Dieb stahl ihr/dein Geld		
A thief stole their money	Ein Dieb stahl ihr Geld			Ein Dieb stahl ihr Geld		
Where is ours	Wo dieses Modell ist			Wo ist unseres		
Where is mine	Wo ist meins			Wo ist meins		
Where is yours	Wo ist deins			Wo ist euers/deins		
Where is theirs	Wo ihnen gehört			Wo ist ihriges		
This is good/this one is good	Das ist gut, / dieses gut ist			Das ist gut, / dieses ist gut.		
Olu is here	Olu hier			Olu ist hier		
Olu was there	Olu war da	5		Olu war da		
Táyé was in the place	Taye war an dem Ort,			Taye war dort		

Táyé arrived yesterday	Taye kam gestern			Taye kam gestern an		
Olú did not arrive yesterday	Olu nicht angekommen gestern			Olu ist nicht gestern angekommen		
Olu had arrived yesterday	Olu gestern angekommen			Olu ist gestern angekommen		
Olu will arrive tomorrow	Olu trifft heute			Olu wird morgen ankommen		
Olu will not arrive tomorrow	Olu nicht morgen ankommen			Olu wird nicht morgen ankommen		
Olu had not arrived	Olu nicht angekommen			Olu war nicht angekommen		
Olu is coming	Olu kommt			Olu kommt		
Olu is coming very soon	Olu kommt sehr bald			Olu kommt sehr bald		
I heard that Olu arrived yesterday.	Ich hörte, dass Olu gestern angekommen			Ich hörte, dass Olu gestern angekommen ist		
Olu is supposed to arrive tomorrow	Olu soll morgen ankommen			Olu soll morgen ankommen		
I want Olu to arrive tomorrow	Ich möchte Olu um anzukommen morgen			Ich möchte, dass Olu morgen ankommt		
I doubt if Olu will arrive tomorrow	Ich bezweifle, dass Olu trifft heute			Ich bezweifle, dass Olu morgen ankommt		
It is good that Olu comes	Es ist gut , dass Olu kommt			Es ist gut, dass Olu kommt		
I command Olu to go immediately	Ich befehle Olu sofort gehen			Ich befehle Olu sofort zugehen		
Olu said I should go out	Olu gesagt, ich soll gehen			Olu hat gesagt, dass ich rausgehen soll		
Olu said you should go out	Olu gesagt, Sie sollten gehen			Olu hat gesagt, dass du/sie rausgehen sollst/sollen		
Did Ibikunle arrive yesterday?	Haben Ibikunle ankommen gestern?			Ist Ibikunle gestern angekommen?		
Is it true that Olu arrived yesterday?	Stimmt es, dass Olu kam gestern ?			Ist es wahr, dass Olu gestern angekommen ist?		
Did Olu arrive yesterday?	Haben Olu ankommen gestern?			Ist Olu gestern angekommen?		

When is he returning?	Wenn er zurückkehrt ?			Wann kommt er wieder?		
How is Olu returning?	Wie wird Olu der Rückkehr ?			Wie kommt Olu wieder?		
Will you ride a bike or drive a car?	Werden Sie ein Fahrrad fahren oder ein Auto fahren ?			Werden sie/Wirst du ein Fahrrad oder ein Auto fahren?		
Where is Olu coming from?	Wo ist Olu aus ?			Woher kommt Olu?		
What has Olu gone to do?	Was hat sich Olu gegangen, um zu tun?			Was würde Olu machen Was ist Olu gegangen, um zu tun?		
Why did you say so?	Warum hast du gesagt?			Warum hast du das gesagt?		
Leave me!	Verlasse mich!			Lass mich in Ruhe!		
You all should leave me alone!	Sie alle sollten mich in Ruhe lassen !			Sie sollten mich alle in Ruhe lassen!		
You, come out here!	Sie , kommen Sie hier !			Komm hier!/Kommen Sie hier		
Children, I said you should come out here	Kinder, sagte ich, sollten Sie hier herauskommen			Kinder, ich habe gesagt, dass ihr hier rauskommen sollt!		
Stop beating the kid	Aufhören zu schlagen das Kind			Hör auf, das Kind zu schlagen!		
Olu is greeting me	Olu ist mir Gruß			Olu grüßt mich		
Olu is greeting you	Olu ist Sie Gruß			Olu grüßt dich/Ihnen/Euch		
Leke is greeting us	Leke wird uns Gruß			Leke grüßt uns		
Leke is greeting them	Leke ist zu grüßen			Leke grüßt sie		
Dada stole Olu’s car	Dada stahlen Olu Auto			Dada hat Olus Auto gestohlen		
Dada is the one who stole Olu’s car	Dada ist derjenige, der Olu Auto gestohlen			Dada ist derjenige, der Olus Auto gestohlen hat		
It is Olú’s car that Dada stole	Es ist Olu Wagen , die Dada gestohlen			Es ist Olus Auto, das Dada gestohlen hat		
It is Olu who Dada stole his car	Es ist Olu , die Dada stahlen sein Auto			Es ist Olu von wem Dada das Auto gestohlen hat		
It was I whose car was stolen by the thief	Ich bin es , dessen Auto wurde von der Dieb gestohlen			Es war ich, der das vom Dieb gestohlen wurde		
He was the one whose car was stolen	Er ist derjenige , dessen Auto gestohlen wurde,			Er ist derjenige, dessen Auto gestohlen wurde,		

We are the ones whose car was stolen	Wir sind derjenige, dessen Auto gestohlen wurde			Wir sind diejenigen, deren Auto gestohlen wurde.		
It was in the market that the thief stole Olu’s car	Es ist auf dem Markt , dass der Dieb stahl Olu Auto			Es war auf den Marktplatz wo der Dieb OlusAuto gestohlen hat		
Is it true that your car was actually stolen?	Stimmt es, dass Sie Ihr Auto tatsächlich gestohlen wurde ?			Stimmt es, dass dein/Ihr Auto tatsächlich gestohlen wurde?		
Surely, he stole the car for the fact that he did not request for it.	Sicherlich , stahl er das Auto für die Tatsache , dass er nicht verlangen, für sie.			Sicherlich stahl er das Auto weil er nicht für es verlangen hat		
Where is Olu who bought the car	Wo ist Olu , die das Auto gekauft haben,			Wo ist Olu, die/der das Auto gekauft hat?		
If you want to buy a car, raise up your hands	Wenn Sie ein Auto kaufen möchten , heben Sie Ihre Hände			Wenn Sie ein Auto kaufen möchten , heben Sie Ihre/heb Deine Hände		
You shouldn’t have done that	Sie sollten das nicht tun sollen			Du hättest das nicht tun sollen		
I arrived when they were eating	Ich komme , wenn sie aßen			Ich bin angekommen als sie aßen		
Once the game started, then heavy rain started	Sobald das Spiel gestartet, so schwere regen begonnen			Sobald das Spiel anfing, begann es heftig zu regnen		
Once the game started, the rain began	Sobald das Spiel gestartet wurde, begann der regen			Sobald das Spiel anfing, begann es zu regnen		
It was not until after the rain stopped, that the game began	Erst nach der regen aufgehört , dass das Spiel begann,			Erst nachdem der Regen aufhörte begann das Spiel		
I entered before the rain started	Ich trat vor der regen begann			Ich bin eingetreten bevor es anfing zu regnen		
If rain starts early, there won’t be famine.	Wenn regen beginnt früh , wird es keine Hungersnot .			Wenn der Regen früh beginnt, wird es keine Hungersnot geben.		
If rain begins early, there will be famine	Wenn regen beginnt früh , wird es Hungersnöte sein			Wenn es früh anfängt zu regnen, wird es Hungersnot geben		
It is fine whether it rains or not	Es ist gut , ob es regnet oder nicht			Es ist gut,ob es regnet oder nicht		

Even if you cry, I will not give you	Auch wenn Sie schreien , werde ich nicht geben Ihnen			Sogar wenn Du weinst/ Sie weinen, werde ich es nicht geben		
Let us work so that we could have money	Lassen Sie uns zusammenarbeiten , damit wir Geld haben			Lass uns uns arbeiten damit wir Geld bekommen		
Let us work to avoid poverty	Lassen Sie uns zusammenarbeiten, um die Armut zu vermeiden			Lass uns arbeiten, um Armut zu vermeiden		
Even though it rained early, there was still famine	Auch wenn es früh regnete , gab es noch Hungersnot			Auch wenn es früh regnete, gab es noch eine Hungersnot		
Even though it did not rain early, there was no famine	Auch wenn es nicht vorzeitig regnen , gab es keine Hungersnot			Obwhol es nicht früh regnete, gab es keine Hungersnot		
I know Olu’s house	Ich weiß, Olu Haus			Ich kenne Olus Haus		
I know Dada’s wife	Ich weiß, Dada Frau			Ich kenne Dadas Frau		
Ibadan is Adelabu’s town	Ibadan ist die Stadt Adelabu			Ibadan ist Adelabus Stadt		
I wore a white cloth	Ich trug ein weißes Tuch			Ich trug ein weißes Tuch		
I wore a red cloth /garment	Ich trug ein rotes Tuch / Bekleidung			Ich trug ein rotes Tuch / eine rote Bekleidung		
I wore a new cloth/garment	Ich trug einen neuen kleiden / Bekleidung			Ich trug ein neues Tuch/eine neue Bekleidung		
I spread a wet cloth	Ich einen nassen kleiden verbreiten			Ich breite nasse Kleidung aus		
I did not spread a dried cloth	Ich habe nicht eine getrocknete kleiden verbreiten			Ich habe nicht trocknete Kleidung ausgebreitet		
I drank a hot tea	Ich trank einen heißen Tee			Ich trank heißen Tee		
I counted a person; two people; three people; four people; five people; six people; seven people; nine people; ten	Ich zählte eine Person ; zwei Menschen; drei Leute; vier Leute; fünf Menschen; sechs Personen ; sieben Personen ; neun Personen ; zehn Personen ; zwanzig Personen			Ich zählte eine Person ; zwei Leute; drei Leute; vier Leute; fünf Menschen; sechs Personen ; sieben Personen ; neun Personen ; zehn Personen ; zwanzig Personen ;		

people; twenty people; thirty people; hundred people	; dreißig Personen ; hundert Menschen			dreißig Personen ; hundert Menschen		
They called first person; second person; third person; fourth person; fifth person; sixth person; seventh person; eighth person; ninth person, tenth person.	Sie nannten ersten Person ; zweite Person ; dritte Person; vierte Person ; fünfte Person ; sechste Person ; siebte Person ; achte Person ; neunte Person , zehnte Person.			Sie riefen erste Person ; zweite Person ; dritte Person; vierte Person ; fünfte Person ; sechste Person ; siebte Person ; achte Person ; neunte Person , zehnte Person.		
Olu went to greet Òjó	Olu ging zu Ojo grüßen			Olu ging um Ojo zu begrüßen		
He went to ask of his welfare	Er ging zu seinem Wohlergehen zu fragen					
They will come to greet us	Sie werden kommen, um uns zu begrüßen			Sie werden kommen, uns zu begrüßen		
They will come to ask of our welfare	Sie werden auch kommen Fragen unserer Wohlfahrt			Sie werden kommen über unserer Wohlfahrt zu fragen		
I picked up the money in the market	Ich nahm das Geld in den Markt			Ich nahm das Geld auf dem Marktplatz auf		
They want to come tomorrow	Sie wollen morgen kommen			Sie wollen morgen kommen		
I saw the money on the ground	Ich sah das Geld auf dem Boden			Ich sah das Geld auf dem Boden		
I saw it on the ground	Ich sah es auf dem Boden			Ich sah es auf dem Boden		
I kept the money in the pocket	Ich hielt das Geld in der Tasche			Ich behielt das Geld in der Tasche		
I kept it in the pocket	Ich hielt es in der Tasche			Ich behielt es in der Tasche		
How did you spend the money	Wie haben Sie das Geld ausgeben			Wie haben Sie das Geld ausgegeben?		
Did you spend the money or keep it	Haben Sie das Geld ausgeben oder halten Sie es			Haben Sie das Geld ausgegeben oder es behalten?		
I can’t remember how I did it again	Ich kann mich nicht erinnern, wie ich es tat wieder			Ich kann mich nicht erinnern, wie ich es wieder tat		

Perhaps it dropped from my pocket	Vielleicht ist es aus der Tasche fallen gelassen			Vielleicht ist es mir aus der Tasche gefallen		
Olu has never been there	Olu war noch nie dort gewesen			Olu ist noch nie dort gewesen		
Olú did not come at all	Olu überhaupt nicht kommen			Olu kam überhapt nicht		
Olu did not come	Olu ist nicht gekommen,			Olu ist nicht gekommen,		