Issue 5 | 2014

SAL

SamaraAltLinguo E-Journal

Published by Samara Alternative Linguistics Project

samaraaltlinguo.com

SAMARAALTLINGUO E-JOURNAL
an interdisciplinary electronic journal

CONTENTS

Applied Linguistics

Investigating language in the machine translation: exploring Yorùbá-English machine translation as a case study
Clement ODOJE
pp 4-11

Language teaching methods to dyslexics acquiring English as a second language
Shini UNNI
pp 12-22

Languages of the World

A communicative and stylistic adaptability of new idioms and idiomatic expressions in Yorùbá literary texts
Dayo AKANMU
pp 23-36

Question-word movement in Tiv
Michael Terhemen ANGITSO
pp 37-52

Applied Linguistics

Investigating language in the machine translation: exploring Yorùbá-English machine translation as a case study

Clement ODOJE
Department of Linguistics and African languages, University of Ibadan, Nigeria
E-mail: lekecment2@gmail.com

ABSTRACT

Computer now translates human language without having any understanding of it. How does this happen? The answer to the question is the concern of this paper. This paper opines that even though machine could only count and match for translation, the deficiencies seen in its translation are not majorly computer's rather the exploration of language acquisition. The paper explores two opposing views on language acquisition and used them to explain the processes of machine translation using Google translator and Ibadan on-going SMT research as example. It concludes that when human can adequately explain human language acquisition then there could be answer to modeling machine to master human language for translation.

KEYWORDS

machine translation
language acquisition
language and environment

1. Introduction

Language is important to translation, at least a translator has to have a good knowledge of languages involved in his translation. But how can we account for language in a machine that translates but does not have any knowledge of natural human language whatsoever or rely on linguistic structures? Manning and Schutze (2000) explain that "some language researchers and many Natural Language Processing (NLP) practitioners are perfectly happy to just work on text without thinking much about the relationship between the mental representation of language and its manifestation in written form. Readers sympathetic with this approach may feel like skipping to the practical sections, but even practically-minded people have to confront the issue of what prior knowledge to try to build into their model, even if this prior knowledge might be clearly different from what might be plausibly hypothesized for the brain". In other words, it is necessary to investigate language representation in the brain while attempting to build a machine that generate, interpret or communicate with language so that an appropriate measure could be adopted for proper explanation of the model adopted in the exercise. Hence, the concern of this paper is to investigate language in the brain via various scholastic discussions and as they relate to machine translation especially statistical approach to Machine Translation (MT). The question to first discuss is how is language acquired?

2. Language acquisition

There have been several explanations and theories on the language acquisition but two prominent opposing views are the Rationalist and Empiricist. It will be better to start this discussion with Sapir's (1912) view which actually led to Sapir-Whorf hypothesis. Sapir links language and culture to the influence of 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 categories environmental influence into two: physical and social factor. To him, any trait of human culture as due solely to the force of physical environment seems to rest on a fallacy. Hence, environment can act directly only on an individual, and in cases where a purely environmental influence is responsible for 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 capable of environmental influence uncombined with influence of another character is doubtful but at least possible hence the smallest environmental influence is either supported or transformed by social forces. On the other hand, 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, physical environment is reflected in language only in so far it has been influenced by social factors.

He explains that language may be influenced in these three ways:

- in regard to its subject matter or content, i.e in regard to the vocabulary;
- in regard to its phonetic system i.e the system of sounds with which it operates in the building of words;
- and in regard to its grammatical form i.e in regard to 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 said:

> *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.*

Hence, other two areas of environmental influence on language i.e. speech sound and grammar are jettisoned in the ways environment influences language.

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

- The influence of environment on language
- The ways language may be influenced
- The main thesis of his claims

I. *The influence of environment on language*: the influence which depends on physical and social factors is incorporated in interest; the term interest is however questioned. Though, it was explained that the physical environment is reflected 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 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 synonymous or antonymous in meaning? How does a child show interest in the heredity and

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? Does it mean that the individual who does not show or have any interest in an item, will have to change its name or the person will forget the name of such item? In short, Sapir did not explain the role of interest in the influence of the environment on language.

No doubt, 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 of the fact that there is no item whether of interest or not of the people in an environment that such people do not have words for. In a situation where there is language contact or foreign item infiltrated into the environment, although it may not have a specific word since it is foreign, then there may be need to look for an equivalence for such item in the language by derivation, description, reduplication, compounding, coinage or borrowing the word from the language of the item or the neighboring languages.

II. *The ways language may be influenced*: Sapir (1912) itemizes three ways by which language can be influenced as shown above but summarised here for convenience; that is: 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 sound and grammar of a language if the child is not in the environment where the language is spoken? Sapir explains that there seems an absolute lack of correlation between physical and social environment and phonetic system, either in their general acoustic aspect or in regard to distribution of particular elements. This is considered as accidental character of a phonetic system. He goes further to explain that the fact that 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 still remains that, though Sapir is focused on the influence of the physical environment without much consideration on the social environment. The social environment provides the child elements of the language for appropriate language acquisition. For example, the child gets the vocabulary, speech sound and grammar from the people around i.e immediate environment. For example, Yorùbá does not have word for *snow* because it does not exist in their environment unlike the Eskimos, Finland, Norway and Sweden who have many equivalent words for it. The best Yorùbá can do is to name it after an item that looks like it *Yìnyín* which is *ice* or better still borrow it as *Sínò.*

III. The main thesis of his claims: the main thesis of Sapir (1912) is that the complexity and rapid change in culture may not necessarily reflect on language. He points out that, cultural element serves the immediate needs of the society and entering more clearly into consciousness which will not only change more rapidly than those of language, but the form itself of culture, giving each element its relative significance, i.e the continually shaping itself anew. Linguistic element 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, conservative tendency makes itself felt more profoundly in the groundwork of language than of culture. He then explains that the consequence of this is that the form of language will in course of time cease to symbolize 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 form and content which is the direct opposite of general view held with respect to the greater conservatism of language in civilized communities than among the primitive people. Then what is the yardstick to measure civilized and primitive peoples'

language if Sapir holds the view that he doubt whether many languages of primitive people have undergone as rapid modification in a corresponding period of time as has the English language?

To this researcher, grouping languages as civilized and primitive is super imposition of a language above others. To Pinker (1995) in Fee (2003) the idea is consider colonialization to assume that there might be important language-based differences among cultures. Every language should and can express thought which proves equality of all languages. Therefore categorizing languages into major, minor, civilized and primitive is basically for political and social reasons not necessarily because a language is more important than other.

3. Sapir-Whorf Hypothesis

Kay & Kempton (1984) explain Sapir-Whorf Hypothesis in relation to thinking. They explain that there are certain thoughts of an individual in one language that cannot be understood by those who live in another language. In other words the way people think is strongly affected by their native languages. Lakoff (1987) in Fee (2003:2) holds the view that Whorf was right in observing that concept that have been made part of the grammar of a language used in thought, not just as object of thought, and that they are used spontaneously, automatically, unconsciously and effortlessly "... I am convinced by Whorf's argument that the way we use concept affects the way we understand experience; concept that are spontaneous, automatic, unconscious are simply going to affect a greater (thoughtless obvious) impact on how we understand everyday life than concept that we merely ponder" (Lakoff 1987:335). Although, with the introduction of Chomskian school of thought, the hypothesis is now believed by most linguists only in the weak sense that language can have some small effect on thought. Kay & Kempton (1984) conclude that the extreme version of the idea that all thought is constrained by language, has been disproved and the opposite extreme that language does not influence thought at all is also widely considered to be false.

To sum it up, Sapir-Whorf Hypothesis is a follow up to the Sapir (1912) claims. There is no dispute in the fact that language is a reflection of conceptual ideas but language is not in any way limited by these concepts. For example, the conceptual idea of gender is not observable in the grammar of a language like Yorùbá compared with languages like German, French, Italian, Spanish etc. But does that mean Yorùbá are not thinking of gender issues? To this researcher, Sapir-Whorf Hypothesis helps in translation in the sense that it help to redefine translation better because translation is not word for word translation or mere text as Catford (1965:1) famous definition of translation put it neither is it limited to stylistic translation like Osundare (1995) but representing concepts of a language in another language. For example a word may represent many things in a language depending on the audience. Consider Olamide a Nigerian musician who just released an album titled *Single-Kó dúró sókè*

Ṣe Kó dúró sókè, kó wálẹ̀
Q should stay at up, should come down
Should it stay up or come down

The meaning of the above extract has different meanings to different audiences based on the concept that they have in mind. Among the public bus drivers and commuters, it is a special form of greetings. This greeting is done by raising up of two hands to salute a superior colleague and ask the superior colleague if the hands are to remain up or come down in order to show absolute loyalty to the recipient of the greetings. So the question the person doing the greeting asks is the above extract. However, in some quarters, among some adults of the same language, the extract above could mean sexuality. The kind of pounding, upward and

downward movements in the act warrant the question in the above extract. Hence, a translator is supposed to have the basic understanding of the language but more importantly the understanding of the target audience for proper and appropriate translation. This is why target audience should be identified during translation. Apart from perception and conceptual knowledge in translation, Sapir-Whorf Hypothesis is basically not universal in it approach to language acquisition explanations.

4. Innateness of language

The hypothesis that claims universality of language is innate hypothesis. Chomsky's position is to correct the claims of behaviorism and structuralism on the language acquisition explanations. His view is channeled via what he called *Poverty of the Stimulus*.

Putnam (1967) explains that Innate Hypothesis hypothesizes that the human brain is 'programmed' at birth in some quite specific and structured aspects of human natural language of which this programming are spelled out in 'Explanatory Models in Linguistics'. He explains further that speakers have 'built in' function which assigns weights to the grammars G_1 G_2, G_3,... in a certain class of transformational grammars whereby is not the class of all possible transformational grammars; rather all the members of have some quite strong similarities. These similarities are technically called Universality of Language or Principles. However, Chomsky (1987) explains the concept of Principles from the perspective of differences in languages which is technically referred to as Parameters that:

> *Languages, of course, differ; English is not Japanese. But it seems that languages differ only in their lexical choices and in selection of certain options that are not fully determined by the fixed principles of our biological endowment. Thus in every language, verbs take objects; but the object may follow the verb, as in English, or precede it, as in Japanese. This option holds not only for verb phrases, but for all phrases. Thus English has prepositions, while Japanese has postpositions. Japanese in many ways seems a mirror image of English, and seems superficially to differ in many other respects as well. But the systems are cast to the same mold.*

In other worlds all human beings have the capacity to acquire language and all languages employ the same principle though the application of these principles differ from one language to another as observed in Chomsky's example in the above extract. Chomsky (1987, 2002, 2006) explains that human language faculty is like any other biological endowment which needs nurturing from the environment. This idea of Chomsky has been criticized (see Putnam 1967, Jackendoff 2012). One crucial problem in accounting for the origins of human language is how an ability to acquire language could have arisen when there was no language around to learn. But this problem is not unique to the evolution of language. It arises in trying to account for any communicative capacity in any organism (Fitch 2011). So this is not a definitive argument against the language capacity having evolved incrementally.

A person's language organ is what Mendívil-Giró (2009:15) refers to as internal language (i-language) which is traditionally or conventionally regarded as Universal Grammar (UG) while external language (E-language) consists of a population of i-languages that allow their possessors to communicate with each other; they are regarded as social institutions (Saussure1916). This paper objects to Mendívil-Giró's (2009) claim that "the i-language of a person who speaks French and that of a person who speaks Russian are historically different" in the sense that since UG is the innate capacity of language for a man to acquire a language, there is no basis for the UG or i-language as it were of a person who speaks French to be different from a person who speaks Russian. If this is to go by, it implies that a person who speaks French and a person who speaks Russian have different language in their DNA. Then, if

language is part of the component of human brain, then all human being must have the same i-language but that raise the question about the diversity in e-language which is one of the major criticism against the claims of the proponent of innateness of language. Fitch (2009:8) explains further that we are born with a language acquisition 'instinct' but not language *per se* meaning that i-language is the language acquisition 'instinct' while e-language is the language per se. Fitch's position is totally a deviant from the explanation of Mendivil-Giro on language innateness.

5. Computer and Human Natural Language

Computer is human invention and it uses formal language to operate, in other words, natural language is meant for humans not computers. How then does a computer translates human natural language? The answer is modelling. Bar-Hillel (1953) in his explanation of the challenges of Machine Translation in the early days of MT is of the view that machine can only count and match. In other words, programming a computer is to tell a computer what to do which is represented in numbers (binary, hex, oct) and expressed in formulae. It is left for humans to write the programme which allows computer to model human natural language.

Two ways of programming computer to model human language is suggested, rule machine learning and statistical machine learning[1]. Rule learning requires specific rules or patterns from which computers learn from and generalize. It implies that a specific rule of operation is given to the computer in order to perform a task. Since what human beings learn is not everyday expressions rather the parametric variations or language specific rule (e-language) as discussed above (fine rule to generate infinite sentences). the concern of a programmer is to convert the syntactic rule of a language or more to formal language for a computer to learn from so as to model the language(s) and if necessary use the language(s) for translation. This process is called Rule Based Approach to Machine Translation. Examples are: Akinola (2009), Odoje (2010), Eludiora et.al (2011) were the Syntactic rules of English and Yorùbá are used to generate sentences and translate the languages bi-directionally. The challenge of this learning is that, sometimes some of the rules cannot be expressed in formal/mathematical formula in which case such rule will not me modelled for computer to learn from hence there might to be desired result(s) in such occasion. A way out of this is to present the computer with raw data so as to figure out the patterns and form the formula which is the statistical machine learning.

Statistical 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 reason from externally supplied instances to produce general hypothesis which then make predictions about 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 word, supervised learning involves the intervention of humans in the process of learning which may be very expensive most especially when there are myriad specifications 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 sample data or the environment. The output states are defined implicitly by the specific learning algorithm

[1] Dr Tunde Adegbola brought up the idea during one of our interactions in an ongoing research before further readings were made.

used and built in constraints.

Relating this to MT therefore, it shows the melting point of Chomsky's and Sapir's idea on the role of environment in the language acquisition. If there is an i-language without the input of the environment to activate it, then there will be a defection or total lack of language acquisition. Statistical Machine Translation (SMT) relies much on the environment (in this case corpus) learn from so as to translate. In one of the personal interaction with Professor Koehn, he explains that to start to build an SMT a corpus with nothing less than a hundred thousand or more will be required. This pose a great challenge to most African languages because since most of the equivalent translation that are available are not digitalised like the European languages hence most SMT are focused on the mainstream languages like, English, Russian, Chinese etc. Of recent, African cross border languages like Arabic and Ki-Swahili are getting involved. Google just lunch Yorùbá as one of the languages available on its translator. The translator translates every simple sentence if the source language is English but translating the same translated sentence (now Yorùbá) to the Source language will generate an unknown sentence. Consider the example below:

	English sentence	Yorùbá translated sentence	Human translation of English sentence	English equivalence of Yorùbá sentence
1	I love Nigeria	Mo ni ife Nigeria	Mi ní ìfẹ́ Nàìjíríà	I love Nigeria
2	He loves Nigeria	O si fẹràn Nigeria	Ó fẹ́ràn Nììjíríà	You like Nigeria
3	I want to eat	Mo fe lati je	Mo fẹ́ jẹun	I want to eat
4	Bola and Tolu eat yam	Bola ati Tolu je isu	Bọ́lá àti Tolú jẹ ìṣu	Honors and perseverance income
5	He kicked the bucket	O si gba awọn garawa	Ó ta téru nípàá	You mumb punch
6	He is to be blamed	Oun ni o ni lati sima	Oun ni ẹ̀bi yẹ/ẹ di ẹ̀bi lé e lórí	You blame it on
7	He knew his wife and she gave birth to twins	O mọ iyawo rẹ ati o fun ibi lati ibeji	Ó mọ ìyàwó rẹ̀, ó sì bí ìbejì	You know your wife, and twin
8	He made all things beautiful	O si se ohun gbogbo lẹwa	Ó ṣe ohun gbogbo dára dára	You do what all good
9	It was possible courtesy his effort	O ti see se iteriba re akitiyan	Ó ṣeé ṣe látàrí/nípasẹ̀ akitiyan rẹ̀	It is possible duento your efforts

You will observe from the translations above that some translations could not represent the meaning of the source language; example is 5, 6, 7, 8 and 9. The reason could be lack of equivalent translated material as mentioned above which could be related to environment in the language acquisition theories.

In other to meet up with the challenges of inadequate corpus; Ibadan SMT adopted D.O Fagunwa's books and other literary text that have equivalent translations. Most of these literary texts are written in old orthography which inform Google's translator not having appropriate diacritics which affect its translation. Ibadan SMT has rewritten the texts with old orthography to new orthography for the purpose of the research. This helps the computer in

modelling Yorùbá language statistically; the research is still on-going.

6. Conclusion

It is true that computer now translates human language without understanding it; this is made possible only through modelling. To achieve remarkable impart in this direction, human beings have to device a means to understand language acquisition in the human mind/brain so as to model that in a computer for optimum achievement. There is also the need to have enough digital translated equivalent corpora to serve as environment for computer to learn from. When these two basic requirement is met, then there will be the need to meet the complexity and dynamism of language which is restricted to human and may not necessary be achieved by machine which impede the development of Fully Automated Machine Translation (FAMT) which is the aim of MT from the beginning.

References:

1. Awofolu, O. 2002. *The Making of A Yoruba-English Machine Translator*. St Mary's City: St Mary College of Maryland.
2. Bar-Hillel Y. 1953. "Some Linguistic Problems Connected with Machine Translation". *Philosophy of Science*, vol.20 , pp 217-225.
3. Bar-Hillel Yehoshua. 1953. "Some Linguistic Problem Connected with Machine Translation". *Philosophy of Science*, Vol 20, pp 217-225.
4. Catford J.C. 1965. *A Linguistic Theory of Translation*. London. Oxford University Press.
5. Chomsky Noam. 1987. "Language, Language Development and Reading", in *Reading Instruction Journal* New Jersey.
6. Chomsky Noam. 2005. "Three Factors in Language Design Linguistic" in *Linguistic Inquiry* Vol 26, No1.
7. Chomsky Noam. 2006. *Language and Mind*. Cambridge. Cambridge university press.
8. Eludiora, Salawu, Odejobi and Agbeyangi's. 2011. Ife: Machine Translation
9. Fee Margery. 2003. The Sapir-Whorf Hypothesis and the Contemporary Language and Literary Revival among the First Nations in *Canada International Journal of Canadian Studies* 27, Spring 2003.
10. Fitch Tecumseh. 2009. "Prolegomena to a Future Science of Biolinguistics" in *Biolinguistics* Vol 3, No 4 pp 283- 320.
11. Koehn Philip. 2010. *Statistical Machine Translation*, Cambridge. Cambridge University Press
12. Lopez A. 2008. Statistical machine translation. ACM Comput. Surv., 40, 3, Article 8 (August 2008), 49 pages DOI: 10.1145/1380584.1380586. Nigeria.
13. Manning Christopher & Schutze Hinrich. 2000. *Foundation of Statistical Natural Language Processing*. London. The MIT press.
14. Mendílvil-Giró Jose-Luis. 2009. "What is a Language from Biolinguistic Point of View?" *Biolinguistics* Vol 3, No 4 pp 283- 320.
15. Osundare Niyi. 1995. "Caliban's Gamble: The Stylistic Repercussion of Writing African Literature in English in Owolabi Kola" (ed) *Language in Nigeria*. Ibadan. Group Publishers.
16. Putnam Hilary. 1967. "The 'Innateness Hypothesis' and Explanatory Model in Linguistics" in *Sythese*, Vol 17, No1 Springer http://www.jstor.org/stable/20114532, accessed:08/01/2013.
17. Sapir Edward. 1912. "Language and Environment" in *The American Anthropologist* Vol 14 pp 226-242.

Issue 5 | 2014

SAL

SamaraAltLinguo E-Journal

Published by Samara Alternative Linguistics Project

samaraaltlinguo.com

SAMARAALTLINGUO E-JOURNAL
an interdisciplinary electronic journal

CONTENTS

Applied Linguistics

Investigating language in the machine translation: exploring Yorùbá-English machine translation as a case study
Clement ODOJE
pp 4-11

Language teaching methods to dyslexics acquiring English as a second language
Shini UNNI
pp 12-22

Languages of the World

A communicative and stylistic adaptability of new idioms and idiomatic expressions in Yorùbá literary texts
Dayo AKANMU
pp 23-36

Question-word movement in Tiv
Michael Terhemen ANGITSO
pp 37-52

Applied Linguistics

Investigating language in the machine translation: exploring Yorùbá-English machine translation as a case study

Clement ODOJE
Department of Linguistics and African languages, University of Ibadan, Nigeria
E-mail: lekecment2@gmail.com

ABSTRACT

Computer now translates human language without having any understanding of it. How does this happen? The answer to the question is the concern of this paper. This paper opines that even though machine could only count and match for translation, the deficiencies seen in its translation are not majorly computer's rather the exploration of language acquisition. The paper explores two opposing views on language acquisition and used them to explain the processes of machine translation using Google translator and Ibadan on-going SMT research as example. It concludes that when human can adequately explain human language acquisition then there could be answer to modeling machine to master human language for translation.

KEYWORDS

machine translation
language acquisition
language and environment

1. Introduction

Language is important to translation, at least a translator has to have a good knowledge of languages involved in his translation. But how can we account for language in a machine that translates but does not have any knowledge of natural human language whatsoever or rely on linguistic structures? Manning and Schutze (2000) explain that "some language researchers and many Natural Language Processing (NLP) practitioners are perfectly happy to just work on text without thinking much about the relationship between the mental representation of language and its manifestation in written form. Readers sympathetic with this approach may feel like skipping to the practical sections, but even practically-minded people have to confront the issue of what prior knowledge to try to build into their model, even if this prior knowledge might be clearly different from what might be plausibly hypothesized for the brain". In other words, it is necessary to investigate language representation in the brain while attempting to build a machine that generate, interpret or communicate with language so that an appropriate measure could be adopted for proper explanation of the model adopted in the exercise. Hence, the concern of this paper is to investigate language in the brain via various scholastic discussions and as they relate to machine translation especially statistical approach to Machine Translation (MT). The question to first discuss is how is language acquired?

2. Language acquisition

There have been several explanations and theories on the language acquisition but two prominent opposing views are the Rationalist and Empiricist. It will be better to start this discussion with Sapir's (1912) view which actually led to Sapir-Whorf hypothesis. Sapir links language and culture to the influence of 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 categories environmental influence into two: physical and social factor. To him, any trait of human culture as due solely to the force of physical environment seems to rest on a fallacy. Hence, environment can act directly only on an individual, and in cases where a purely environmental influence is responsible for 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 capable of environmental influence uncombined with influence of another character is doubtful but at least possible hence the smallest environmental influence is either supported or transformed by social forces. On the other hand, 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, physical environment is reflected in language only in so far it has been influenced by social factors.

He explains that language may be influenced in these three ways:

- in regard to its subject matter or content, i.e in regard to the vocabulary;
- in regard to its phonetic system i.e the system of sounds with which it operates in the building of words;
- and in regard to its grammatical form i.e in regard to 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 said:

> *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.*

Hence, other two areas of environmental influence on language i.e. speech sound and grammar are jettisoned in the ways environment influences language.

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

- The influence of environment on language
- The ways language may be influenced
- The main thesis of his claims

I. *The influence of environment on language*: the influence which depends on physical and social factors is incorporated in interest; the term interest is however questioned. Though, it was explained that the physical environment is reflected 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 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 synonymous or antonymous in meaning? How does a child show interest in the heredity and

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? Does it mean that the individual who does not show or have any interest in an item, will have to change its name or the person will forget the name of such item? In short, Sapir did not explain the role of interest in the influence of the environment on language.

No doubt, 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 of the fact that there is no item whether of interest or not of the people in an environment that such people do not have words for. In a situation where there is language contact or foreign item infiltrated into the environment, although it may not have a specific word since it is foreign, then there may be need to look for an equivalence for such item in the language by derivation, description, reduplication, compounding, coinage or borrowing the word from the language of the item or the neighboring languages.

II. *The ways language may be influenced*: Sapir (1912) itemizes three ways by which language can be influenced as shown above but summarised here for convenience; that is: 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 sound and grammar of a language if the child is not in the environment where the language is spoken? Sapir explains that there seems an absolute lack of correlation between physical and social environment and phonetic system, either in their general acoustic aspect or in regard to distribution of particular elements. This is considered as accidental character of a phonetic system. He goes further to explain that the fact that 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 still remains that, though Sapir is focused on the influence of the physical environment without much consideration on the social environment. The social environment provides the child elements of the language for appropriate language acquisition. For example, the child gets the vocabulary, speech sound and grammar from the people around i.e immediate environment. For example, Yorùbá does not have word for *snow* because it does not exist in their environment unlike the Eskimos, Finland, Norway and Sweden who have many equivalent words for it. The best Yorùbá can do is to name it after an item that looks like it *Yìnyín* which is *ice* or better still borrow it as *Sínò.*

III. The main thesis of his claims: the main thesis of Sapir (1912) is that the complexity and rapid change in culture may not necessarily reflect on language. He points out that, cultural element serves the immediate needs of the society and entering more clearly into consciousness which will not only change more rapidly than those of language, but the form itself of culture, giving each element its relative significance, i.e the continually shaping itself anew. Linguistic element 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, conservative tendency makes itself felt more profoundly in the groundwork of language than of culture. He then explains that the consequence of this is that the form of language will in course of time cease to symbolize 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 form and content which is the direct opposite of general view held with respect to the greater conservatism of language in civilized communities than among the primitive people. Then what is the yardstick to measure civilized and primitive peoples'

language if Sapir holds the view that he doubt whether many languages of primitive people have undergone as rapid modification in a corresponding period of time as has the English language?

To this researcher, grouping languages as civilized and primitive is super imposition of a language above others. To Pinker (1995) in Fee (2003) the idea is consider colonialization to assume that there might be important language-based differences among cultures. Every language should and can express thought which proves equality of all languages. Therefore categorizing languages into major, minor, civilized and primitive is basically for political and social reasons not necessarily because a language is more important than other.

3. Sapir-Whorf Hypothesis

Kay & Kempton (1984) explain Sapir-Whorf Hypothesis in relation to thinking. They explain that there are certain thoughts of an individual in one language that cannot be understood by those who live in another language. In other words the way people think is strongly affected by their native languages. Lakoff (1987) in Fee (2003:2) holds the view that Whorf was right in observing that concept that have been made part of the grammar of a language used in thought, not just as object of thought, and that they are used spontaneously, automatically, unconsciously and effortlessly "... I am convinced by Whorf's argument that the way we use concept affects the way we understand experience; concept that are spontaneous, automatic, unconscious are simply going to affect a greater (thoughtless obvious) impact on how we understand everyday life than concept that we merely ponder" (Lakoff 1987:335). Although, with the introduction of Chomskian school of thought, the hypothesis is now believed by most linguists only in the weak sense that language can have some small effect on thought. Kay & Kempton (1984) conclude that the extreme version of the idea that all thought is constrained by language, has been disproved and the opposite extreme that language does not influence thought at all is also widely considered to be false.

To sum it up, Sapir-Whorf Hypothesis is a follow up to the Sapir (1912) claims. There is no dispute in the fact that language is a reflection of conceptual ideas but language is not in any way limited by these concepts. For example, the conceptual idea of gender is not observable in the grammar of a language like Yorùbá compared with languages like German, French, Italian, Spanish etc. But does that mean Yorùbá are not thinking of gender issues? To this researcher, Sapir-Whorf Hypothesis helps in translation in the sense that it help to redefine translation better because translation is not word for word translation or mere text as Catford (1965:1) famous definition of translation put it neither is it limited to stylistic translation like Osundare (1995) but representing concepts of a language in another language. For example a word may represent many things in a language depending on the audience. Consider Olamide a Nigerian musician who just released an album titled *Single-Kó dúró sókè*

Ṣe Kó dúró sókè, kó wálę̀
Q should stay at up, should come down
Should it stay up or come down

The meaning of the above extract has different meanings to different audiences based on the concept that they have in mind. Among the public bus drivers and commuters, it is a special form of greetings. This greeting is done by raising up of two hands to salute a superior colleague and ask the superior colleague if the hands are to remain up or come down in order to show absolute loyalty to the recipient of the greetings. So the question the person doing the greeting asks is the above extract. However, in some quarters, among some adults of the same language, the extract above could mean sexuality. The kind of pounding, upward and

downward movements in the act warrant the question in the above extract. Hence, a translator is supposed to have the basic understanding of the language but more importantly the understanding of the target audience for proper and appropriate translation. This is why target audience should be identified during translation. Apart from perception and conceptual knowledge in translation, Sapir-Whorf Hypothesis is basically not universal in it approach to language acquisition explanations.

4. Innateness of language

The hypothesis that claims universality of language is innate hypothesis. Chomsky's position is to correct the claims of behaviorism and structuralism on the language acquisition explanations. His view is channeled via what he called *Poverty of the Stimulus*.

Putnam (1967) explains that Innate Hypothesis hypothesizes that the human brain is 'programmed' at birth in some quite specific and structured aspects of human natural language of which this programming are spelled out in 'Explanatory Models in Linguistics'. He explains further that speakers have 'built in' function which assigns weights to the grammars G_1 G_2, G_3,... in a certain class of transformational grammars whereby is not the class of all possible transformational grammars; rather all the members of have some quite strong similarities. These similarities are technically called Universality of Language or Principles. However, Chomsky (1987) explains the concept of Principles from the perspective of differences in languages which is technically referred to as Parameters that:

> *Languages, of course, differ; English is not Japanese. But it seems that languages differ only in their lexical choices and in selection of certain options that are not fully determined by the fixed principles of our biological endowment. Thus in every language, verbs take objects; but the object may follow the verb, as in English, or precede it, as in Japanese. This option holds not only for verb phrases, but for all phrases. Thus English has prepositions, while Japanese has postpositions. Japanese in many ways seems a mirror image of English, and seems superficially to differ in many other respects as well. But the systems are cast to the same mold.*

In other worlds all human beings have the capacity to acquire language and all languages employ the same principle though the application of these principles differ from one language to another as observed in Chomsky's example in the above extract. Chomsky (1987, 2002, 2006) explains that human language faculty is like any other biological endowment which needs nurturing from the environment. This idea of Chomsky has been criticized (see Putnam 1967, Jackendoff 2012). One crucial problem in accounting for the origins of human language is how an ability to acquire language could have arisen when there was no language around to learn. But this problem is not unique to the evolution of language. It arises in trying to account for any communicative capacity in any organism (Fitch 2011). So this is not a definitive argument against the language capacity having evolved incrementally.

A person's language organ is what Mendívil-Giró (2009:15) refers to as internal language (i-language) which is traditionally or conventionally regarded as Universal Grammar (UG) while external language (E-language) consists of a population of i-languages that allow their possessors to communicate with each other; they are regarded as social institutions (Saussure1916). This paper objects to Mendívil-Giró's (2009) claim that "the i-language of a person who speaks French and that of a person who speaks Russian are historically different" in the sense that since UG is the innate capacity of language for a man to acquire a language, there is no basis for the UG or i-language as it were of a person who speaks French to be different from a person who speaks Russian. If this is to go by, it implies that a person who speaks French and a person who speaks Russian have different language in their DNA. Then, if

language is part of the component of human brain, then all human being must have the same i-language but that raise the question about the diversity in e-language which is one of the major criticism against the claims of the proponent of innateness of language. Fitch (2009:8) explains further that we are born with a language acquisition 'instinct' but not language *per se* meaning that i-language is the language acquisition 'instinct' while e-language is the language per se. Fitch's position is totally a deviant from the explanation of Mendivil-Giro on language innateness.

5. Computer and Human Natural Language

Computer is human invention and it uses formal language to operate, in other words, natural language is meant for humans not computers. How then does a computer translates human natural language? The answer is modelling. Bar-Hillel (1953) in his explanation of the challenges of Machine Translation in the early days of MT is of the view that machine can only count and match. In other words, programming a computer is to tell a computer what to do which is represented in numbers (binary, hex, oct) and expressed in formulae. It is left for humans to write the programme which allows computer to model human natural language.

Two ways of programming computer to model human language is suggested, rule machine learning and statistical machine learning[1]. Rule learning requires specific rules or patterns from which computers learn from and generalize. It implies that a specific rule of operation is given to the computer in order to perform a task. Since what human beings learn is not everyday expressions rather the parametric variations or language specific rule (e-language) as discussed above (fine rule to generate infinite sentences). the concern of a programmer is to convert the syntactic rule of a language or more to formal language for a computer to learn from so as to model the language(s) and if necessary use the language(s) for translation. This process is called Rule Based Approach to Machine Translation. Examples are: Akinola (2009), Odoje (2010), Eludiora et.al (2011) were the Syntactic rules of English and Yorùbá are used to generate sentences and translate the languages bi-directionally. The challenge of this learning is that, sometimes some of the rules cannot be expressed in formal/mathematical formula in which case such rule will not me modelled for computer to learn from hence there might to be desired result(s) in such occasion. A way out of this is to present the computer with raw data so as to figure out the patterns and form the formula which is the statistical machine learning.

Statistical 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 reason from externally supplied instances to produce general hypothesis which then make predictions about 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 word, supervised learning involves the intervention of humans in the process of learning which may be very expensive most especially when there are myriad specifications 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 sample data or the environment. The output states are defined implicitly by the specific learning algorithm

[1] Dr Tunde Adegbola brought up the idea during one of our interactions in an ongoing research before further readings were made.

used and built in constraints.

Relating this to MT therefore, it shows the melting point of Chomsky's and Sapir's idea on the role of environment in the language acquisition. If there is an i-language without the input of the environment to activate it, then there will be a defection or total lack of language acquisition. Statistical Machine Translation (SMT) relies much on the environment (in this case corpus) learn from so as to translate. In one of the personal interaction with Professor Koehn, he explains that to start to build an SMT a corpus with nothing less than a hundred thousand or more will be required. This pose a great challenge to most African languages because since most of the equivalent translation that are available are not digitalised like the European languages hence most SMT are focused on the mainstream languages like, English, Russian, Chinese etc. Of recent, African cross border languages like Arabic and Ki-Swahili are getting involved. Google just lunch Yorùbá as one of the languages available on its translator. The translator translates every simple sentence if the source language is English but translating the same translated sentence (now Yorùbá) to the Source language will generate an unknown sentence. Consider the example below:

	English sentence	Yorùbá translated sentence	Human translation of English sentence	English equivalence of Yorùbá sentence
1	I love Nigeria	Mo ni ife Nigeria	Mi ní ìfẹ́ Nàìjíríà	I love Nigeria
2	He loves Nigeria	O si fẹràn Nigeria	Ó fẹ́ràn Nììjíríà	You like Nigeria
3	I want to eat	Mo fe lati je	Mo fẹ́ jẹun	I want to eat
4	Bola and Tolu eat yam	Bola ati Tolu je isu	Bọ́lá àti Tolú jẹ ìṣu	Honors and perseverance income
5	He kicked the bucket	O si gba awọn garawa	Ó ta téru nípàá	You mumb punch
6	He is to be blamed	Oun ni o ni lati sima	Oun ni ẹ̀bi yẹ/ẹ di ẹ̀bi lé e lórí	You blame it on
7	He knew his wife and she gave birth to twins	O mọ iyawo rẹ ati o fun ibi lati ibeji	Ó mọ ìyàwó rẹ̀, ó sì bí ìbejì	You know your wife, and twin
8	He made all things beautiful	O si se ohun gbogbo lẹwa	Ó ṣe ohun gbogbo dára dára	You do what all good
9	It was possible courtesy his effort	O ti see se iteriba re akitiyan	Ó ṣeé ṣe látàrí/nípasẹ̀ akitiyan rẹ̀	It is possible duento your efforts

You will observe from the translations above that some translations could not represent the meaning of the source language; example is 5, 6, 7, 8 and 9. The reason could be lack of equivalent translated material as mentioned above which could be related to environment in the language acquisition theories.

In other to meet up with the challenges of inadequate corpus; Ibadan SMT adopted D.O Fagunwa's books and other literary text that have equivalent translations. Most of these literary texts are written in old orthography which inform Google's translator not having appropriate diacritics which affect its translation. Ibadan SMT has rewritten the texts with old orthography to new orthography for the purpose of the research. This helps the computer in

modelling Yorùbá language statistically; the research is still on-going.

6. Conclusion

It is true that computer now translates human language without understanding it; this is made possible only through modelling. To achieve remarkable impart in this direction, human beings have to device a means to understand language acquisition in the human mind/brain so as to model that in a computer for optimum achievement. There is also the need to have enough digital translated equivalent corpora to serve as environment for computer to learn from. When these two basic requirement is met, then there will be the need to meet the complexity and dynamism of language which is restricted to human and may not necessary be achieved by machine which impede the development of Fully Automated Machine Translation (FAMT) which is the aim of MT from the beginning.

References:

1. Awofolu, O. 2002. *The Making of A Yoruba-English Machine Translator*. St Mary's City: St Mary College of Maryland.
2. Bar-Hillel Y. 1953. "Some Linguistic Problems Connected with Machine Translation". *Philosophy of Science*, vol.20 , pp 217-225.
3. Bar-Hillel Yehoshua. 1953. "Some Linguistic Problem Connected with Machine Translation". *Philosophy of Science*, Vol 20, pp 217-225.
4. Catford J.C. 1965. *A Linguistic Theory of Translation*. London. Oxford University Press.
5. Chomsky Noam. 1987. "Language, Language Development and Reading", in *Reading Instruction Journal* New Jersey.
6. Chomsky Noam. 2005. "Three Factors in Language Design Linguistic" in *Linguistic Inquiry* Vol 26, No1.
7. Chomsky Noam. 2006. *Language and Mind*. Cambridge. Cambridge university press.
8. Eludiora, Salawu, Odejobi and Agbeyangi's. 2011. Ife: Machine Translation
9. Fee Margery. 2003. The Sapir-Whorf Hypothesis and the Contemporary Language and Literary Revival among the First Nations in *Canada International Journal of Canadian Studies* 27, Spring 2003.
10. Fitch Tecumseh. 2009. "Prolegomena to a Future Science of Biolinguistics" in *Biolinguistics* Vol 3, No 4 pp 283- 320.
11. Koehn Philip. 2010. *Statistical Machine Translation*, Cambridge. Cambridge University Press
12. Lopez A. 2008. Statistical machine translation. ACM Comput. Surv., 40, 3, Article 8 (August 2008), 49 pages DOI: 10.1145/1380584.1380586. Nigeria.
13. Manning Christopher & Schutze Hinrich. 2000. *Foundation of Statistical Natural Language Processing*. London. The MIT press.
14. Mendílvil-Giró Jose-Luis. 2009. "What is a Language from Biolinguistic Point of View?" *Biolinguistics* Vol 3, No 4 pp 283- 320.
15. Osundare Niyi. 1995. "Caliban's Gamble: The Stylistic Repercussion of Writing African Literature in English in Owolabi Kola" (ed) *Language in Nigeria*. Ibadan. Group Publishers.
16. Putnam Hilary. 1967. "The 'Innateness Hypothesis' and Explanatory Model in Linguistics" in *Sythese*, Vol 17, No1 Springer http://www.jstor.org/stable/20114532, accessed:08/01/2013.
17. Sapir Edward. 1912. "Language and Environment" in *The American Anthropologist* Vol 14 pp 226-242.