THE RELATION BETWEEN LANGUAGE AND INFORMATION TECHNOLOGY (IT)

Egenti, Martha Chidimma
Department of Linguistics
Nnamdi Azikiwe University Awka
egentinma@gmail.com

&

Dereck-M Akachukwu Orji
Department of Linguistics
Nnamdi Azikiwe University Awka

Abstract

In the growth of technological advancements, there is need to highlight some of the technological tools that enhance academic research in linguistics at different levels of grammar given the availability of software that improves linguistic analysis. This study is simply geared towards bringing to limelight some of these linguistics tools and their uses as a way of creating awareness. Also, the relation between language and information technology (IT) shows that language is a sole carrier of volumes of information while IT is a veritable tool through which data are analyzed and interpreted.

Keywords: language, information, technology

1. Introduction

Today, technological advancements have permeated every nook and cranny of the world and have turned it into a global village. People from all walks of life; including professionals, students, and the general population are confronted by unprecedented volumes of information, the vast bulk of which is stored as unstructured text.

Bird, Klein and Loper (2008:14), notes that in 2003, it was estimated that the annual production of books amounted to 8Terabytes. (A Terabyte is 1,000 Gigabytes, i.e., equivalent to 1,000 pickup trucks filled with books.) It would take a human being about five years to read the new scientific material that is produced every 24 hours. Although these estimates are based on printed materials, increasingly the information is also available electronically. Indeed, there has been an explosion of text and multimedia content on the World Wide Web. For many people, a large and growing fraction of work and leisure time is spent navigating and accessing this universe of information. Language on the other hand, is the sole carrier of this bulk of information which is embodied in the several language technologies whether it is spoken or written data. We shall in the subsequent sections look at the language and its richness, give an overview of information technology and language technology and specific areas of linguistics in which computers are used for linguistic analysis.

2. Language and its Richness

According to Bird, Klein and Loper (2008:14), Language is rich. It is the chief manifestation of human intelligence. Through language we express basic needs and lofty aspirations, technical know-how and flights of fantasy. Ideas are shared over great separations of distance and time. The study of language is part of many disciplines outside of linguistics, including translation, literary criticism, philosophy, anthropology and psychology. Many less obvious disciplines investigate language use, such as law, hermeneutics, forensics, telephony, pedagogy, archaeology, cryptanalysis and speech pathology. Each applies distinct methodologies to gather observations, develop theories and test hypotheses. Yet all serve to deepen our understanding of language and of the intellect that is manifested in language. The importance of language to science and the arts is matched in significance by the cultural treasure embodied in language.

No doubt, as Bird et al (2008) further point out, each of the world's over 7,000 human languages is rich in unique respects, in its oral histories and creation legends, down to its grammatical constructions and its very words and their nuances of meaning. Threatened remnant cultures have words to distinguish plant subspecies according to therapeutic uses that are unknown to science. Language is indispensable both in arts & humanities and in science and engineering as well. In arts and humanities, programming is carried out to manage language data, explore linguistic models, and test empirical claims; likewise, in science and engineering, language serves as a source of interesting problems in data modeling, data mining, and knowledge discovery. It is also important to note that as Anusuya and Katti

(2010) in distinguishing between human brain and the computer assert that Language is the key linkage which makes both brain and computer to work efficiently.

Language is dynamic, as Bird et al (2008) note evolves over time as they come into contact with each other and they provide a unique window onto human pre-history. Technological change gives rise to new words like blog and new morphemes like e- and cyber-. For its breathtaking complexity and diversity, human language is as a colorful tapestry stretching through time and space. The next section will look at information technology and language technology.

3.0 Information Technology and Language technology

We shall first and foremost look at what information technology entails before delving into language technology. Information Computer Technology also called Information technology (IT) according to Urua (2004:5) involves the processing of information, the storage of such information, the distribution and assessment of such information with the use of the computer, modern technologies and digital devices. Information/data can be organized in a friendly way using computer. IT is the current research tool utilized in virtually all walks of life. Its resources include; internet and World Wide Web (WWW). The internet is a flexible computer-based global system with many interconnected computer networks, linking thousands of computers to enable them share information. The web consists of programs running on many computers and permits a user to find and display multimedia (multimodal) and documents (documents with a combination of text, photographs, graphics, audio and video). Similarly, Emeka-Nwobia (2008) also explains ICT and IT differently. According to him, whereas ICT involves the storage, management, exchange, transmission, manipulation, retrieval and assessing of information in its various forms through technological device. These devices may be computers, telephones, networks, electronic databases, CD ROMs, laser discs, video cameras, or fax machines; on the other hand, IT refers to the applications of technology to the acquisition, storage, organization, manipulation, assessing, referring and dissemination of information.

On the other hand, according to Rashel (2011) Language technology which is sometimes referred to as human language technology includes: computational methods, computer programs and electronic devices that are specialized for analyzing, producing or modifying texts and speech. She further explains that these systems must be based on few knowledge of human language. Similarly, Uszkoreit (1997) defines information technologies as specialized for dealing with the most complex information medium in our world: human language. Therefore, these technologies are also often subsumed under the term Human Language Technology. Human language occurs in spoken and written form. Whereas speech is the oldest and most natural mode of language communication, complex information and most of human knowledge is maintained and transmitted in written texts. Speech and text technologies process or produce language in these two modes of realization. But language also has aspects that are shared between speech and text such as dictionaries, most of grammar and the meaning of sentences. Thus, large parts of language technology cannot be subsumed under speech and text technologies. Among those are technologies that link language to knowledge.

Furthermore, language technologies can also help people communicate with each other. Much older than communication problems between human beings and machines are those between people with different mother tongues. One of the original aims of language technology has always been fully automatic translation between human languages. One of the major goals as Rashel (2011) points out is to simplify and improve the communication system between people and computer. From bitter experience scientists have realized that they are still far away from achieving the ambitious goal of translating unrestricted texts. Nevertheless, they have been able to create software systems that simplify the work of human translators and clearly improve their productivity. Less than perfect automatic translations can also be of great help to information seekers who have to search through large amounts of texts in foreign languages.

According to Uszkoreit (1997) in our communication we mix language with other modes of communication and other information media. We combine speech with gesture and facial expressions. Digital texts are combined with pictures and sounds. Movies may contain language and spoken and written form. Thus speech and text technologies overlap and interact with many other technologies that facilitate processing of multimodal communication and multimedia documents. Nevertheless, Uszkoreit also points out that language is the fabric of the web, according to him, the rapid growth of the internet/www and the emergence of the information society poses exciting new challenges to language technology. However, he asserts that although, the new media combine text, graphics, sound and movies, the whole world of multimedia information can only be structured, indexed and navigated through language.

3.1 The Relation between Language and Information Technology

Computational Linguistics is largely an applied field, concerned with practical problems. There are as many applications as there are reasons for computers to process or produce language: for example, in situations where humans are unavailable, too expensive, too slow, or busy doing tasks that humans are better at than machines.

Some current application areas include translating texts, especially business and technical texts from one language to another (Machine Translation), finding relevant documents in large collections of text (Information Retrieval), and answering questions about a subject area (expert systems with natural language interfaces).

The most relevant language technologies are summarized by Uszkoreit (1997) below:

- **Speech recognition**

Spoken language is recognized and transformed in into text as in dictation systems, into commands as in robot control systems, or into some other internal representation.

- **Speech synthesis**

Utterances in spoken language are produced from text (text-to-speech systems) or from internal representations of words or sentences (concept-to-speech systems). Similarily, Urua (2004), explains that speech synthesis enable the computers to scan printed text and speak the words aloud. This tool is useful for the blind and visually challenged people.

- **Text categorization**

This technology assigns texts to categories. Texts may belong to more than one category, categories may contain other categories. Filtering is a special case of categorization with just two categories.

- **Text Summarization**

The most relevant portions of a text are extracted as a summary. The task depends on the needed lengths of the summaries. Summarization is harder if the summary has to be specific to a certain query.

- **Text Indexing**

As a precondition for document retrieval, texts are stored in an indexed database. Usually a text is indexed for all word forms or – after lemmatization – for all lemmas. Sometimes indexing is combined with categorization and summarization.

- **Text Retrieval**

Texts are retrieved from a database that best match a given query or document. The candidate documents are ordered with respect to their expected relevance. Indexing, categorization, summarization and retrieval are often subsumed under the term information retrieval.

- **Information Extraction**

Relevant information pieces of information are discovered and marked for extraction. The meaning of the texts is analyzed in detail, answering specific questions about text content. The extracted pieces can be: the topic, named entities such as company, place or person names, simple relations such as prices, destinations, functions etc. or complex relations describing accidents, company mergers or football matches. Other kinds of text such as medical case histories that are in well-bounded domain systems will extract information and put it into databases for statistical analyses.

- **Data Fusion and Text Data Mining**

Extracted pieces of information from several sources are combined in one database. Previously undetected relationships may be discovered

3.2 Computers in Specific Areas of Linguistics and Resources of IT

Rogers (1998) highlights core areas of linguistics in which computers are used for linguistic analysis.

Phonetics and Phonology

Computers helps to improve the way phonetics is taught but to be able to do it with the computer, the sounds of a particular language must be digitized which can be fed into any computer for linguistic analysis. There is also some available software like audacity software where these sounds can be edited to remove background noise.

Rogers also notes that phonemic transcription can be taught using a computer by application of the appropriate software. Also, computer offers the opportunities for teaching anatomy needed in phonetics. The computer provides not only the access to the data, but also the tools for the analysis. CD-ROMs are used in storing large amounts of data. Programs used for phonetics and phonology among others include; PRAAT, E-LAN, CAPTAC, Phthong etc.

Morphology

The computer can be used to perform tasks such as dividing a word into morphemes, listing other allomorphs, identifying port-manteau morphs. It can also give the total of words in a given text etc.

Syntax

Tree diagram for sentences can be easily drawn by using a program that could parse strings. The computer can also convert labeled bracketed strings into trees. E.g SYNTACTICA

Semantics

The computer helps to make fine distinctions of meaning. Program such as Tarski’s World used in teaching logic can be useful in this regard. Others include SEMANTICA, FrameBuilder (allows creation of lexical database with semantic definitions). ELLOGON, AntCONC are concordance tools used to show the keyword-in-context concordance on the basis of frequency, length of word, inclusion in a list, or pattern matching. That is to say different semantic analysis can be performed using them, also the collocations patterns of words and the different meanings words can have in the different occurrences can be analyzed using these software.

Sociolinguistics

Although, combination of some of the above listed software can be used for sociolinguistic analysis. For instance, the E-LAN or PRAAT can be used for annotation of the data. Language variations, innovations and isoglosses in a language can be captured using GAPMAP and other software showing these variations and different dialects on the map.

Language teaching and learning

ICT is learner's friendly. Learners can assess information through the internet with little or no teacher's guidance. Today, we have on-line education program called e-learning whereby people enroll for any degree programme. Academic materials can also be accessed in libraries of universities in the world. The virtual library provides this service and grant researchers access to books, journals, magazines, course curricula, university programmes, e.t.c.

Other areas such as Psycholinguistics, language documentation etc. are also applied to computational method in the analysis of data. A lot of languages have been described and preserved computationally using IT resources. IT such as Shoebox and Microsoft Excel and Access can also be used for building database for dictionary making thereby building large corpora for a particular language.

3.3 Some Software for linguistic analysis

There are lots of software that is relevant for different aspects of research in linguistics. They can also be combined in the different stages of data analysis. For instance, AUDACITY is used for recording and processing recordings. ELAN/PRAAT is used for transcription/annotation. PRAAT is used for acoustic/auditory analysis and for generating output; it is basically used by linguists for speech analysis, NOTEPAD++ is used for preparing scripts and formatting data tables. While EXCEL is used for quick or basic plots, spot checks, data manipulation etc.; R is used for statistics and advanced plots. ELLOGON/ANTCONC is used for key word-in-context or concordances, etc. The following Software is also used for dialect analysis in sociolinguistics and especially for dialect variations studies. It includes:

Lo4:http://www.let.rug.nl/kleiweg/Lo4/
Gabmap: http://www.gabmap.nl
VDM (Visual Dialectometry): www.dialectometry.com
LingPy: http://lingPY.org. This website used for historical linguistic studies.
Most of the above software are free online and they can be easily downloaded for use for different linguistic analysis.

4. Summary and Conclusion

The foregoing no doubt presents IT resources as an indispensable tool that helps in solving language problems. Today, it is possible to manage very large corpora using the CD ROMs, MP3, FLASH DRIVE etc. Data can be accessed easily in the internet. The different software used in collection of text, annotation and plotting of graphs such as E-LAN, AUDACITY, Ant-CONC, R-Brul, PRAAT e.t.c. are all found in the internet using the Google search engine. Most of these software are free online software and can be downloaded alongside their tutorials which are also available online. Using IT for linguistic analysis or for academic purposes not only saves time, it makes work a lot easier and more fun. Following Anusuya and Katti (2010) this study also notes that "…language is the key linkage which makes both brain and computer to work efficiently".

References

Anusuya M.A. & S. K. Katti .2010. Superficial Analysis and Differences between the Human Brain and the Computer. *International Journal of Computer science and Network Security*, vol. 10 N0 7.

Emeka-Nwobia Ngozi. 2008. Information and communication technology (ICT) in Mbah and Mbah (Ed.), *History of Linguistics and Communication*, a festschrift in honor of Professor P.A. Nwachukwu. Nsukka-Enugu: Paschal Communication, pp. 358-369.

Bird S. E. Klein & Loper E. 2008. Introduction to Natural language processing. Retrieved from http://nltk.org/book/

Rashel M. 2011. Introducing Language Technology and Computational Linguistics in Bangladesh. International Journal of English Linguistics. Vol. 1, No. 1. www.ccsenet.org/ijel.

Rogers Henry. 2008. Education in Using Computers in Linguistics: a practical guide in John Lawler and Helen Aristar Dry (Ed.). London and New York: Routledge.

Uszkoreit H. 1997. Language Technology – A first Overview in Survey of the state of the art in Human language technology by Ronald A-Cole, Joseph Mariam, Uszkoreit Hans, Zaenen Annie and Zue Victor (Ed.). Retrieved from http://www.dfki.de/~hansu/LT.pdf. on 15/2/15.