PSYCHOLOGY AND EDUCATION (2024) 61(12): 735-745
ISSN: 1553-6939

Computational Linguistics and Natural Language Processing

Si Bachir Rachid[1], Ajou Menad[2]

[1]Ahmad Zabana University of Relizane (Algeria).

[2]Djilali Liabes University, Sidi Belabbes (Algeria).

The Author's E-mail: si.rachid80@gmail.com[1], menad.adjou@dl.univ-sba.dz[2]

Received: 06/2024 **Published: 12/2024**

Abstract:

Computational linguistics is an interdisciplinary field that combines computer science and linguistics, focusing on the computational processing of linguistic phenomena through the collaboration of linguists and computer scientists. This scientific interplay is evident in the way in which computers deal with linguistic phenomena in a similar way to the human mind, making the automated processing of natural languages the central focus of interest and care. Researchers have developed specific programs and systems that use computer data to study various linguistic questions and phenomena.

In this research, we will explore the basic concepts of computational linguistics as well as the procedures for automated processing of language.

Keywords: computational linguistics, automated language processing, linguistic programming, language computation methods.

Introduction:

Linguists have invested in technological advances in the field of procedural applied linguistics; they have developed tools for studying language and analysing its phonetic, morphological and syntactic levels. They have benefited from laboratory techniques of linguistic analysis and innovations in artificial intelligence, carrying out experimental research to uncover its secrets and depths. They have also used the results of other sciences related to language, such as linguistics, artificial intelligence, computer science, etc.

All this has led to the emergence of "computational linguistics" as a modern interdisciplinary research field that combines computer science and linguistics, with different terminologies such as computational linguistics, mechanical linguistics, media linguistics and routine linguistics all referring to a single concept.

From this point, computational programming emerges as a distinct new field in modern linguistic research, focusing on the computer-based processing of linguistic phenomena. This scientific

convergence is evident in the interaction of the computer with linguistic phenomena, similar to the interaction of the human mind with them.

The automated processing of natural languages is at the centre of attention, where specialised researchers have developed specific programs and systems that use computer data to study various linguistic issues and phenomena at their phonetic, morphological, syntactic, rhetorical and machine translation levels, as well as language teaching and the automated analysis of written and spoken texts, among others, with the aim of achieving practical goals.

1. Computational linguistics

Several modern linguistic trends have emerged in the light of contemporary technological developments, including computational linguistics and cognitive linguistics. The former - computational (or algorithmic) linguistics (Linguistique Computationnelle) - is one of the applied branches. In addition to this term, there are several others, such as Natural Language Processing (NLP) and Human Language Technologies. Despite the differences in terminology among specialists, all these terms revolve around a common theme, which is evident in the following definitions. They see it as "the science of training computer systems to understand human language and simulate human intelligence"[1].

Abdul Rahman Al-Haj Saleh notes that the first beginnings were marked by "the interest of specialists in the mechanical processing of language and its conceptualisation in symbolic form, starting with Noam Chomsky and the theory of transformational grammar"[2]. In essence, it is "the science of making computer systems understand human language and simulate human intelligence"[3]. This is based on a "theoretical conception that imagines the computer as a human mind and attempts to study the cognitive and psychological processes that the human mind uses to produce, understand and perceive language"[4].

Thus, this science - computational linguistics - focuses on using computer data to study various linguistic issues, as well as using computers to serve language and develop methods for learning it. This is in line with the innovation of computer and information software and the invention of machines that rival human thought, mind and intelligence.

Nabil Ali added, attributing the reasons for this convergence to:

1. The enormous development in the field of linguistics, with its aspects subjected to mathematical, logical and statistical treatment.

[1]- Mohsen Rachwan, Introduction to Language Computing, in a collective book Introduction to Arabic Language Computing, King Abdullah bin Abdulaziz International Center for Arabic Language Service, 1st ed., 1441 AH - 2019 CE, Riyadh, Saudi Arabia, p. 17).
[2]- Abdel Rahman Al-Haj Saleh, Patterns of Computational Linguistic Formulation and Modern Khalilian Theory, Journal of the Algerian Academy of the Arabic Language, No. 6, Algeria, 2007, pp. 10-11.
[3]- Mohsen Rachwan,Introduction to Language Computing, p. 17.
[4]- Jallabli Soumeya,Applied Linguistics: Its Concept and Fields, Al-Athar Journal, No. 29, September 2017, p. 131.

2. The emergence of information theory, which provided the mathematical basis for measuring the amount of information.

3. The progress made in computer science (programming languages, automata theory).

4. Progress in mathematical statistics and the infiltration of its methods into linguistic analysis.

5. The appearance of supercomputers and the expansion of artificial intelligence systems, of which linguistic processing mechanisms are one of the most important components. This has allowed the development of automated language processing systems.

6. The first appearance of expert systems that mimic the tasks of experts.

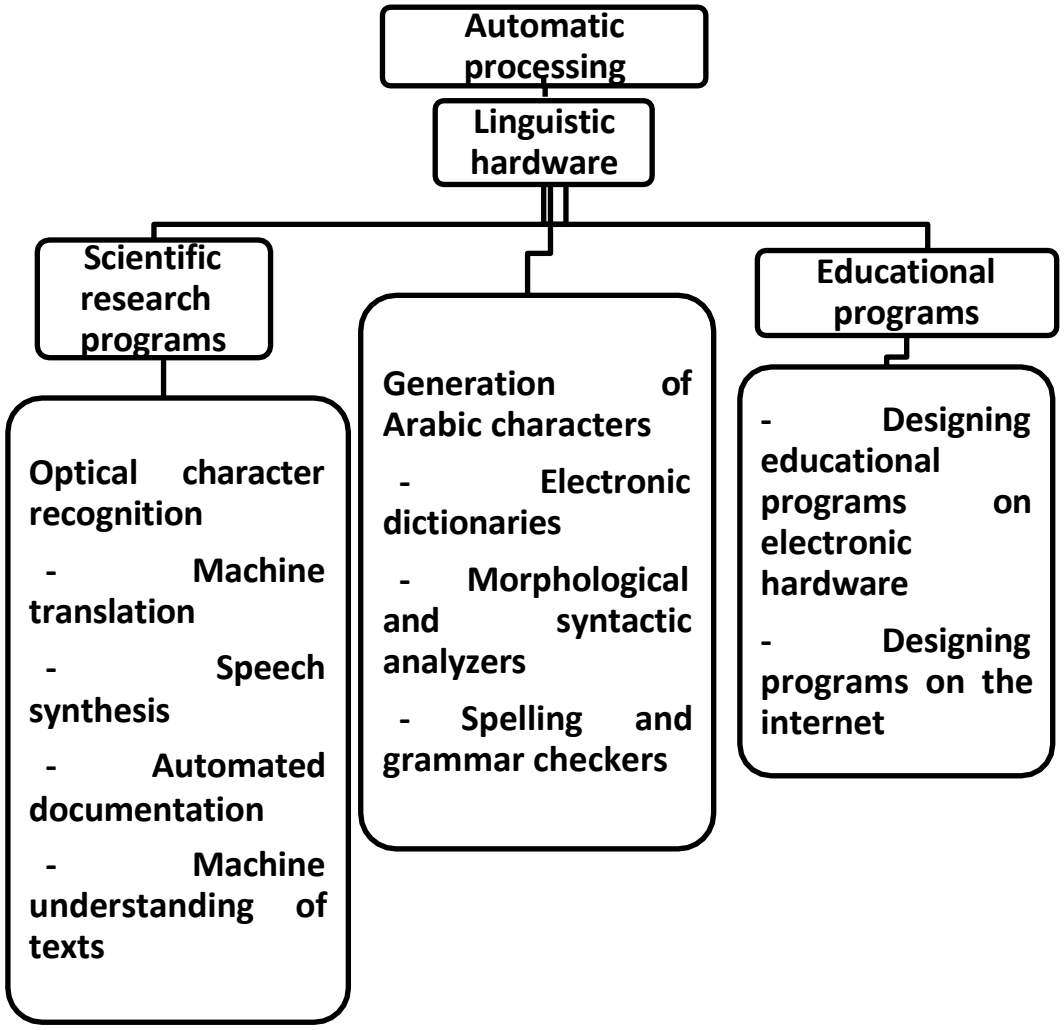
7. The entry of computer applications into the humanities and the widespread use of computers as a means of teaching and learning in general, and language teaching in particular.

The series of factors that have strengthened the link between language and computers can be summarised by the development of the language-computer-application triad[5].

Automated language processing is seen as a cognitive field that transcends the limits of innate human intelligence and ventures into the realm of artificial intelligence, where scientific efficiency reaches its peak. Language is thus given a procedural horizon that distances it from abstract processing, relying on the scientific competence of the learner to draw from the constantly evolving technological context produced by the technological age to which linguistic theory has responded.

Some associate it with a group of sciences, defining it as: “the simulation of the human mind in the theoretical and practical understanding of linguistic phenomena”[6]. This field of knowledge combines linguistics, artificial intelligence, computer science, mathematics and logic, with the aim of transferring human intelligence to computational intelligence*, enabling it to analyse the linguistic system through automated multi-level analysis in the shortest possible time. This is illustrated in the following diagram[7], which shows the fields of automated natural language processing and their interrelationships.

[5]- See: Nabil Ali, Arabic Language and Computer, Kuwait Institute for Arabization, (n.d.), 1988, pp. 114-116.
[6]- Brahimi Boudaoud,Arabic Language Computing in Light of Informational Renewal: "The Linguistic Treasure Project," Professor Abdel Rahman Haj Saleh as a Model, Knowledge Bridges Journal, Hassiba Ben Bouali University, Chlef, Algeria, Vol. 05, No. 01, March 2019, p. 17.
* Informatics: Refers to computer science (/computer science/ l’informatique).
[7]- Khalifa Al-Misawi, Linguistic Terminology and Conceptual Foundations, Dar Al-Amane, Rabat, 1st ed., 2013, pp. 30-31.



- Figure 1: Areas of research in computational processing of natural languages ([8])

It is clear from this description that the computer's simulation of human language is a process that involves many sciences, and its ability to process linguistic data with its various phenomena. Nabil Ali explains this scientific interplay between linguistics and computer science as follows "Language and computers converge for a fundamental and simple reason, which is that language embodies what is essential in man, namely his mental activity in all its manifestations. At the same time, the computer aims to simulate some human functions and mental capacities, taking human considerations as the primary axis for the design of its systems, applications and operational requirements"[9].

Computers operate on a similar basis to the human mind, but no one can claim that human and artificial intelligence are identical. What is certain is that there are many similarities between them.

[8]- See: Mohammed Mohammed Al-Hanash, Arabic Linguistic Engineering: A Quick Read on Arabic Linguistic Engineering or An Approach to Simulating the Arabic Brain Linguistically, Ajman Network for Science and Technology Journal, Vol. 10, No. 3, 2005, pp. 212-214.

[9]- Nabil Ali, Arabic Language and Computer, Arabization Foundation, Kuwait (n.d.), 1988, p. 114.

The computer has surpassed all previous scientific achievements; indeed, all subsequent scientific and civilisational achievements would not have been possible without the computer's ability to solve many of their complex problems and overcome their obstacles. The computer is a machine that imitates human functions and mental capacities[10], which means that it can simulate the way the human mind works in several areas. Therefore, we find that the Central Processing Unit (CPU)* in a computer, which is the electronic brain invented by the human mind, is merely a tool for performing operations with ease, precision and great speed compared to the time it takes the human mind to process them. This happens after its memory has been charged and equipped with a set of algorithms translated into machine language instructions and commands.

Thus, computational linguistics aims to "explain how the human mind works in its interaction with language in terms of knowledge, acquisition and use"[11], using the computer and its many advantages and capabilities, such as storing, programming and retrieving information at the appropriate time. This has had a significant impact and effectiveness in language education, as modern technological means have been used in teaching, leading to the emergence of new fields such as language computing, e-learning, machine translation and automated grammar checking, among others.

Computational linguistics enables computers to interact with linguistic phenomena in a way similar to human cognition. It rests on three main pillars:

1. Text processing techniques: Examples include machine translation, automatic summarisation and text mining.

2. Speech processing techniques: Examples include automatic speech recognition and the conversion of written text into spoken language.

3. Image processing techniques: An example is optical character recognition (OCR)[12].

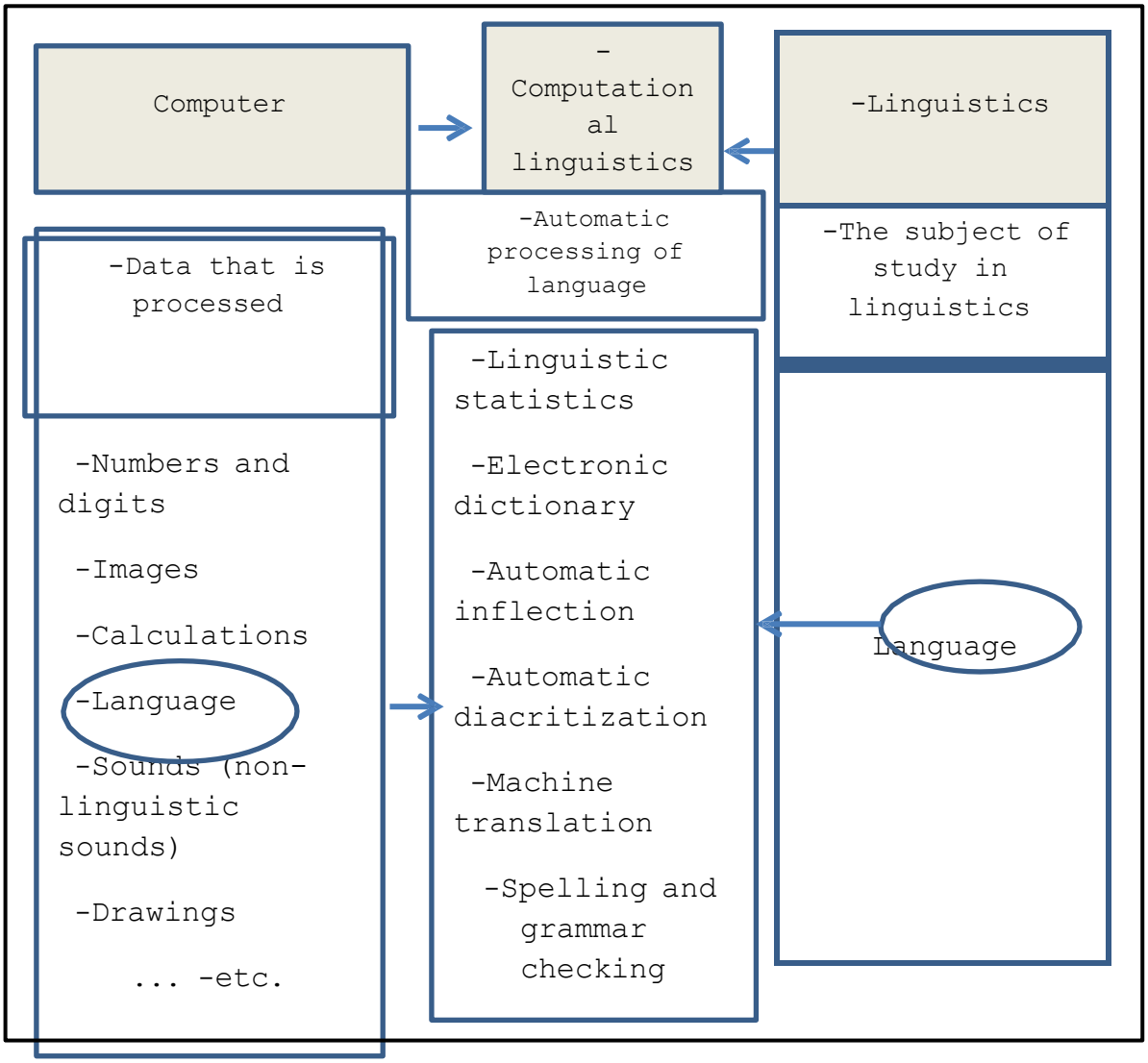
Thus, computational linguistics is the result of the automatic processing of natural language using programming languages such as Python and a range of hardware including computers, scanners, sound processing devices and more. The results of this process are a collection of programs and applications that can be installed on computers, digital tablets and even mobile phones to support language learning and use. It is about translating and studying linguistic phenomena and what the human mind produces into practical computer applications.

Therefore, users of application software directly related to linguistic studies recognise that there is a manifestation of mental activity; these programs are designed to simulate human cognitive functions and abilities.

[10]- Samir Sharif, Linguistics: Field, Function, and Method, Modern Book World, Jordan, 2nd ed., 2008, p. 528.
*The processor is the essential part of the computer's architecture, performing automatic processing of data and information after receiving commands to produce results, and it conducts arithmetic and logical operations.
[11]- Al-Yubi, Belkacem, Computational Linguistics: Its Concept, Developments, and Applications, (Envisioning New Horizons to Serve the Arabic Language and Culture), Meknasa Journal, 1999, No. 12, p. 44.
[12]- Mahmoud Mustafa Khalil, Assigning Verbs to Pronouns in Light of Computational Linguistics, the previous reference, p. 6.

Accordingly, computational linguistics is an interdisciplinary science that integrates linguistics with computer science and other fields such as mathematical logic and artificial intelligence. It aims at the scientific study of natural languages, including morphology, syntax and more, based on advanced and developed systems and programs using computers. The field brings together linguists and computer scientists.

The fields of computational linguistics have diversified, especially in language teaching, leading to the development of programmes and the design of educational websites for this purpose. The aim of all this is to improve the level of education and to use computational linguistics to facilitate language teaching and its application at all levels, including phonetics, morphology, syntax and lexicon. Therefore, it is essential to use computational linguistics in teaching Arabic to non-native speakers, especially by enabling artificial intelligence to capture the unique aspects of this language. The following diagram illustrates the main areas of focus:



- Figure 2: Areas of computational linguistics.

2. Procedural Aspects of Language Computing

Computational linguistics is "part of artificial intelligence, a specialised scientific field that aims to program computers to think, solve problems and make decisions in a way that mimics human capabilities"[13].

In light of this modern linguistic transformation, linguistic computing has become one of the most important contemporary issues occupying the minds of linguistic researchers who seek to design and build linguistic computing programs. These programs use modern technologies to solve linguistic and phonetic problems, as well as automatic processing of language levels, thereby contributing to the solution of language problems and their application in language education and development. In this way, the field of linguistics is upgraded to meet the demands of the era of development and technology.

From this perspective, it is essential for all parties involved in this science - computational linguistics - including linguists, computer scientists, programmers and others, to be familiar with the latest developments in computer science, which is an important link between linguistics and computer science.

In this context, computation relies on a set of inputs aimed at achieving accurate results. It depends on a systematic framework consisting of "a collection of interconnected parts that interact with the environment and with each other to achieve a specific goal by accepting inputs and producing outputs through organised transformation"[14].

Thus, language computing enables the computer to deal with linguistic phenomena in a similar way to the human mind. This is achieved through the collaboration of linguists and computer scientists. The linguist provides all the details of the linguistic phenomenon to be computed, while the computer scientist (programmer) translates these details, after modelling and structuring them with the help of the linguist, into algorithms that ultimately lead to computational applications.

2-1 Modelling and Formalisation

Modelling is a necessary stage for the success of automated processing and requires the collaboration of linguists and computer scientists. It is defined as "a principle or technique that allows the researcher to build a model of a phenomenon or behaviour by quantifying the variables or the explanatory factors for each of these variables"[15]. It is a scientific approach that allows the understanding of complex and intricate systems by creating a model that serves as a formal structure that virtually reproduces reality. The formal part includes a set of calculations and mathematical

[13]- Emad Al-Sabbagh, Information Systems: Their Nature and Components, Dar Al-Thaqafa Library for Publishing and Distribution, Amman, 1st ed., 2000, p. 13.
[14]- Abdel Rahman Mohammed Ta'ma,Constructive Epistemology of Sciences: A Comparative Approach to the Contemporary Linguistic Model, Arabic Language Journal, High Council for the Arabic Language, No. 38, (pp. 13-66), pp. 35-36.
[15]- See: Mustafa Gelfan et al., Generative Linguistics: From the Pre-Standard Model to the Adnoy Program: Concepts and Examples, Modern Book World, Irbid, Jordan, 1st ed., 2010, pp. 19-20.

equations, as well as a collection of abstract rules and logical symbols, where each symbol has a unique, fixed definition determined by the rules that govern and control its operation[16].

Thus, formalisation represents scientific theories within a formal system that allows a clear distinction between linguistic expressions and accepted rules of proof[17]. The aim of formalisation is to produce a logical or mathematical representation of a given phenomenon or process, ensuring that "the process of analysis is completely unambiguous and that the model used in the analysis is verifiable in terms of the mechanisms of operation of its components". Therefore, the formal system must unambiguously define its values according to regulated rules of use, and the operations to which this symbol is subjected must also be precisely defined by the axioms on which the formal system is based[18].

Consequently, formalisation involves subjecting language rules to mathematical symbols in order to study any linguistic structure for automated programming. It establishes laws to control its units in terms of form and function, or, in other words, it works to create a logical or mathematical representation of a phenomenon or a specific process, thereby building the computational system for that phenomenon; that is, "the transition of linguistic material from a linguistic model to a mathematical algorithmic model suitable for automated programming, through the regulation of appropriate linguistic theories to constrain constants and regulate their variables"[19].

2-2 - Algorithms

The term "algorithm" is associated with the Arab mathematician Al-Khwarizmi. The stage of preparing algorithms is a preliminary step in computer programming and is based on what the linguist has achieved in the modelling stage. An algorithm is defined as "a sequential set of defined and countable operations necessary to accomplish a task or solve a particular problem and obtain a correct result"[20]. In other words, it is a method of thinking and analysis that must be followed in order to correctly write commands in a programming language, since it is "based on three basic principles: sequence, selection and repetition, just as the human brain does when it tries to arrange and organise its thoughts in a sequence"[21].

Thus, the preparation of algorithms is essential for analysis, as they analyse the data describing the linguistic phenomenon. They are the ground rules on which automatic language processing and the construction and design of programmes are based. They provide a precise description of the steps required to create a computer application program.

[16]- See: Encyclopédie Universalis, corpus 9, Européenne (unions - Gauguin, Publisher in Paris, France, 2002, p. 634).
[17]- The same reference, pp. 228.
[18]- Sihem Moussaoui et al., Automated Processing of Natural Languages: Arabic Language as a Model, Alfa for Documents Publishing and Distribution, Amman, Jordan, 1st ed., 2021, pp. 29-30.
[19]- Suleiman Aida Al-Muhammadi, Algorithms and Principles of Programming, Fortorn University, Damar, Faculty of Engineering, 2018, p. 29.
[20]- Sihem Moussaoui et al., Automated Processing of Natural Languages: Arabic Language as a Model, p. 40.
[21]- Suleiman Aida Al-Muhammadi, Algorithms and Principles of Programming, Fortorn University, Damar, Faculty of Engineering, 2018, p. 30.

It is certain that the programmer relies on analytical methods and thinking to formulate algorithms, which makes it easier to write programming language commands correctly. Algorithms can be formulated in several ways, the most important of which are:[22]

- Writing them in natural language, such as Arabic or English.

- Formulating them graphically using flowcharts (graphical representation)*.

- Formulating them using a specific symbolic language.

In addition, when designing algorithms specific to linguistic phenomena, we rely on several steps after analysing the phenomenon and understanding and identifying the material related to it. The main steps are:

- Identification of the inputs related to the linguistic phenomenon.

- Identifying the outputs, i.e. the results we want to achieve.

- Determining the processing method, which includes the logical commands and procedures we apply to the inputs to achieve the outputs.

Thus, an algorithm is a series of sequential steps applied to a set of inputs to produce outputs that represent the result and the solution to the problem addressed.

2-3 Databases

Hussein bin Ali Al-Zarayi states that "Computational linguistics, or language computation, is a branch of applied linguistics that focuses on describing and comparing natural languages by building an accurate digital database of linguistic knowledge in all its components and branches using various computer sciences. It uses software programs to link natural language databases with artificial intelligence language rules to enable the retrieval and invocation of stored linguistic data"[23]. Thus, computational linguistics is based on incorporating the rules of natural language, with all its laws, into accurate digital databases on a computer to represent the linguistic system.

A database is a means of organising and storing information; it is "a collection of related files stored in computer repositories that can be added to or modified"[24]. Accordingly, a database is a set of information and data stored in a standardised manner, without duplication, linked together and separate from the programs and applications that process the data. It can be managed by operations such as retrieving, updating, deleting and adding according to a system called a Database Management System (DBMS).

[22]- The flowchart is a diagrammatic representation of the logic used to solve a problem, consisting of specific shapes and symbols arranged according to the order of solving the problems
* Hussein bin Ali Al-Zara'i,
[23]- Linguistics and Its Cognitive Tools, Arab Publishing House, Beirut, Lebanon, 1st ed., 2016, p. 208.
[24]- Nihad Al-Moussa, Arabic Language: Towards a New Description in Light of Computational Linguistics, Beirut, Arab Institute for Studies and Publishing, 1st ed., 2000, p. 53.

The structure of a database consists of tables that are linked together by either relational, network or hierarchical models. Each table contains a set of columns representing fields, each field having a specific data type, and rows representing records. The database system is therefore fundamental to computing.

The advantages of databases include the speed with which information can be obtained and retrieved when needed, as well as its organisation. Some software programs are responsible for creating these databases, including Access, Oracle, Excel and Database Desktop.

Speech Database

Speech databases form the essential basis for building various computer systems, such as automatic speech recognition systems, text-to-speech systems, speaker recognition, and language and dialect recognition. Speech databases typically consist of audio files (wave files). The richer and more diverse the content of the speech database, the better it contributes to the creation of computational systems with outstanding performance.

2-4 Programming and Programming Language

A programming language is defined as "a set of instructions and commands, written according to a set of rules, that are used to create programs and communicate with the computer to perform the required tasks"[25]. Programming languages are based on a set of concepts and functions. Some of the most common and widely used languages are:

- C++ language:

- Java language: Its main features include support for creating applications with sound and graphics.

- HTML language: Known for its high capabilities, it is used to design web pages.

- PHP language: Used for developing web applications.

- Pascal language: Primarily used in education.

-Python Language: One of the most widely used and advanced languages, used in artificial intelligence.

- Delphi Language: Used in various fields based on database management.

Conclusion

From the above, we can conclude that modern interdisciplinary studies have succeeded in combining two sources: the mind - human intelligence - and the machine - artificial intelligence. Thanks to artificial intelligence, which is based on computing, digitisation and automatic processing of natural

[25]- Tanius Joseph, Informatics, Arabic Language, Literature, and Civilization (Number and Letter), Modern Publishing House, Lebanon, 1st ed., 2012, p. 158.

languages, "computational linguistics" has emerged as a fundamental cornerstone and transformative path in contemporary linguistic studies.

The computational linguistic model represents an interdisciplinary research stream that integrates several fields, including computer science, psychology, linguistics and artificial intelligence. This integration requires that researchers in the field have a solid understanding of its foundations and principles in order to effectively manage computational data and understand the mechanisms of automatic language processing.

The procedural aspects of language processing require the joint efforts of linguists and computer scientists to establish precise scientific rules and foundations, to use advanced equipment and specially designed software, and to create standardised databases that are consistent with and facilitate the application of all scientific research in the field.