

Introducing Language Technology & Computational Linguistics in Bangladesh

Md. Mostafa Rashel

Lecturer, Department of English, Daffodil International University &

M.Phil Researcher, Department of Linguistics, University of Dhaka

4/2, Sobhanbagh, Prince Plaza (3rd Floor), Mirpur Road, Dhanmondi, Dhaka- 1207, Bangladesh

E-mail: md.mostafa_rashel@hotmail.com

Abstract

Information technology should have much to offer linguistics not only through the opportunities offered by large-scale data analysis and the stimulus to develop formal computational models, but through the chance to use language in systems for automatic natural language processing. The paper discusses these possibilities in detail and then examines the actual work that has been done in Bangladesh. It is an evident that this has so far been primary research within a new field of Computational Linguistics (CL) which is largely motivated by the demands and interest of practical processing systems and that information technology has rather little influence on linguistics at large.

Keywords: Language technology, Computational linguistics, Human language, Machine translation, Bangla computing

1. Introduction

Language Technology (LT) is the interface of Computer Science, Linguistics and Psychology. Computational Linguistics (CL) and Natural Language Processing (NLP) are closely related names for this region and overlaps between the disciplines. It concerns how human language is constructed and how a computer may be programmed in order to handle our language. Sometimes Language Technology that also 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. These systems must be based on few knowledge of human language. Therefore language technology defines the engineering branch of computational linguistics. Although the existing language technology systems are far from achieving human ability but they have numerous possible applications.

2. Goals and Scopes of Language Technology

The goal of language technology is to simplify and improve the communication system between people and computer. Language technologies are information technologies that are specialized for dealing with the most complex information medium in the world, which is *human language*. Human language occurs in spoken and written form. Whereas speech is the oldest and most natural model 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, grammar and the meaning of sentences. Thus large parts of language technology cannot be subsumed under speech and text technologies. Among those, technologies are link language to knowledge. It is not possible to find out how language, knowledge and thought are represented in the human brain. Nevertheless, language technology has to create formal representation systems that link language to concepts and tasks in the real world. This provides the interface of the fast growing area of knowledge technologies. In our communication we mix language with other model of communication and other information media (*i.e.* speech, digital text, movies). We combine speech with gesture and facial expressions (H. Uszkoreit, 1997). These combinations can be shown in the following way (*see Fig. 4*).

3. Functions of Language Technology

Language Technology (LT) system is far from achieving human ability but it has numerous possible applications or functions. The goal of language technology is to create software products that have few knowledge of human language. Natural language interfaces enable the user to communicate with the computer in French, English,

German or another human language such as database queries, information retrieval from text, so-called expert systems. Communication with computers using spoken language will have a lasting impact upon the work environment. It is completely new areas of application for information technology that will open up a new platform. However, spoken language needs to be combined with other modes of communication such as pointing with mouse or finger. Language technologies can also help people to communicate with each other. The most relevant language technologies are Speech Recognition, Speech Synthesis, Text Categorization, Text Summarization, Text Indexing, Text Retrieval, Information Extraction, Data Fusion and Text Data Mining, Question Answering, Report Generation, Spoken Dialogue Systems and Translation Technologies (H. Uszkoreit, 1997) (*see Fig. 1*).

4. Computational Linguistics

Computational Linguistics (CL) is a relatively new discipline that lies in the intersection of the fields of linguistics (psychology, logic) and computer science (*see Fig. 2*). Many new hybrid disciplines are involving with computers that require computational expertise as well as a background in another field. The term computational linguistics covers many subfields. It sometimes refers to the use of computers as a tool to understand or implement linguistic theories. So we can definitely say that linguists and computer scientists can gain a better understanding of the scientific and research questions by using computer. On the other hand, the term is sometimes used to refer to working systems or applications where the linguistics knowledge is indeed. In this case, the questions and issues are usually one of software engineering as well as of theory (William & Michael, 1989).

Computational Linguistics (CL) might be considered as a synonym of automatic processing of natural language, since the main task of computational linguistics is just the construction of computer programs to process words and texts in natural language. Actually, Computational linguistics is slightly “more linguistic than computational” (Bolshakov & Gelbukh, 2004:147).

Since the 1970s, it has become apparent how complex language actually is and contemporary computational linguistics makes to use of experts from a number of fields. The investigation and modeling of human language is a truly interdisciplinary endeavor. However, the methods of language technology come from several disciplines such as computer science, computational and theoretical linguistics, mathematics, electrical engineering and psychology (*see Fig. 3*). Machine Translation (MT) has always been a major goal of computational linguistics. This task is very complex such as requiring the identification of parts of speech, an understanding of grammar, an extensive vocabulary and mechanisms for dealing with colloquialisms and slang. Machine translation is far from perfect, but with each year the translations become more accurate and less forced. The Machine Translation (MT) for English-Chinese, English-Arabic, English-French and many other pairs of languages are relatively quite advanced. On the other hand, very little work has been done in developing ‘Bangla Machine Translation’ in the field of automatic translation, parsing and syntax analysis to develop software for translating English-Bangla or Bangla-English vice-versa in Bangladesh. Some work has been done; it’s for the Bangla of West Bengal in India. Here, we are citing some Bangla Machine Translation research work in Bangladesh. A significant part of the development of any Machine Translation system is the creation of lexical resources that the system will use. For accurate and efficient transformation from one language to another the necessity of Machine Translation dictionary is obvious for specific domain. An attempt is made to develop Machine Translation Bangla dictionaries that addressed the organization, contents and details of the information (Ali & Ali, 2002: 272–76). Saha (2005) developed low cost English to Bangla-ANUBAD (Translation) that translates English text into Bangla text with disambiguation. It is used for both the rule-based and transformation-based MT schemes along with three level of parsing. An effort is made to develop a statistical Bangla to English translation engine by using only simple Bangla sentences that contains a subject, an object and a verb (Ashraf & kamal, 2004:545).

Speech Recognition (SR) is another area of computational linguistics, which has seen much public interest. In the early 21st century a number of new speech recognition software suites arrived on the market boasting extensive learning systems and high rates of accuracy. This has led to a renewed interest in speech recognition software by the general public and an accompanying increase in funding and research.

Speech Generation (SG) is a related field of computational linguistics that has seen steady development since the 1980s. Reaching a natural-sounding reading of written text is a very difficult problem but one is that holds enormous potential benefits. For non-sighted users, speech generation software can be critical to enjoying the fruits of the digital age.

Computational linguistics also plays an important role in automated grammar correction systems. An accurate grammar checker requires a sophisticated ability to identify parts of speech and a comprehensive list of grammatical rules and exceptions. While most mainstream grammar checkers still have many problems, they are already becoming indispensable for many in the new generation. Computational linguistics is an exciting field

drawing from a wide range of disciplines. It has addressed many problems but none of these are simple. The dream of a universal translator is to word-perfect speech recognition; the goals of computational linguistics cannot help but evoke a sense of wonder.

5. Objectives of Computational Linguistics

Although the objectives of research in computational Linguistics are widely varied, a primary motivation has always been the development of specific practical systems that involves natural language. Three classes of application, which have been central in the development of computational linguistics, are Machine Translation (MT), Information Retrieval (IR) and Man-machine Interfaces (Grishman, 1994) (*see Fig. 5*).

Work on Machine Translation began in the late 1950s with high hope and little realization of the difficulties involved. Extensive work was done in the early 1960s, but a lack of success and in particular a realization that fully automatic high-quality translation would not be possible without fundamental work on text '*understanding*'. Only a few of the current projects in computational linguistics in the United States are addressed toward machine translation, although there are substantial projects in Europe and Japan (Slocum, 1984, 1985; Tucker 1984). In response to a query, the IR was to extract the relevant text from a corpus and either display the texts in most or use the text to answer the query directly. Because the domains of interest (particularly technical and scientific reports) are quite complex, there was little immediate success in this area, but it led to research in knowledge representation. Automatic information retrieval is now being pursued by a few research groups (Sager, 1978; Hirschman & Sager, 1982; Montgomery, 1983: 55-61)

On the other hand, the need to develop complete '*understanding*' systems has forced computational linguistics to develop areas of research, which had been inadequately explored by the traditional science. Two of these areas are Procedural models of the psychological processes of language understanding and Representation of knowledge (*see Fig. 6*).

6. Application Areas of Computational Linguistics

Computational Linguistics tries to solve many areas but the major areas are Machine Translation, Natural Language Interfaces, Grammar and Style checking, Document processing and Information retrieval and Computer-Assisted Language Learning. Many people with a degree in Computational Linguistics work in research groups in universities, governmental research labs or in large enterprises. For example in Sweden Computational Linguistics works in a research groups at various universities that offer courses in linguistics (like Goteborg or Uppsala) at research labs like SICS (The Swedish Institute of Computer Science) or companies like Telia or IBM.

In addition there are development groups working in commercial products. These range from software houses like Microsoft that employs computational linguistics for their work on Grammar Checkers and Automatic Summarization to the Munich based DailLabs that develops a machine translation system to caterpillar, which employs Computational Linguistics for translations of technical manuals.

In recent years the demand for Computational Linguistics has raised with the increase of Language Technology products in the Internet. Job offers come from developers for improving search engines with linguistic means or facilitating the user interface with longboats. Other is integrating speech recognition with language processing techniques. The main professional organizations in computational linguistics are The Association of Computational Linguistics (ACL), Association for Computers and the Humanities (ACH) and International Association for Machine Translation (IAMT).

7. Prospects and Applications of Computational Linguistics in Bangladesh

Bengali, also called Bangla, is the official language of Bangladesh and the Indian States of West Bengal and Tripura. There are over 230 million native speakers of this language across the world and it has the pride of place as the fifth *most spoken* language in the world (after Mandarin, Spanish, English and Hindustani (Hindi-Urdu) and Bangla is also the second most commonly spoken language in India (after Hindi) (Paul, 2009).

Bangla is the first language of Bangladesh. However, organized efforts in software & computer based content and software system localization in Bangla are not very visible in the country. It is obvious that before any content can be generated or any application developed, some basic standards for encoding the language must be developed. The first attempt to use Bangla in computing was made in the early 1980s with Bangla font development in the Windows environment. Commercial vendors led these efforts. It was in 1986 when Bangla language first entered the computer system through '*Shahid Lipi*'. It was a breakthrough. The introduction of '*Bijoy*' Bangla software also added a new dimension to the Bangla computing initiative. The main problem at that period was the compatibility issue of Bangla language in different areas. Bangla was not usable as a general

language on every system as there was no unique way to represent Bangla. In the late 1990s Unicode shed new light on the issue and the process of Bangla computing began to take a new shape in the country (Islam, 2009).

The open source movement has some impact on Bangla in computing. In 1998, J. Ahmed (wiki.mozilla.org/L10n:Teams:bn-BD), a software developer in Bangladesh, first solved the Bangla issue in computing and started a process of Bangla version of Linux. In the late 1990s, a voluntary group named Ankur (www.ankurbangla.org) started Bangla open source software like Linux, OpenOffice.org, Gaim etc. Another voluntary organization, Ekushey (ekushey.org), started developing open source Unicode fonts and a Bangla input system (*i.e.* determining how Bangla fonts can be arranged using the existing keyboard). In 2004, the Bangladesh Computer Council (BCC) took the initiative from the government side and came up with a national Bangla keyboard mapping and a collation sequence. Around this time, The Center for Research on Bangla Language Processing (CRBLP) of BRAC University is currently conducting research projects that deal with Bangla language processing. At present the research team is working on Bangla Document authoring, Information Retrieval (*i.e.*, Spelling checker, Search Engine etc.), Optical Character Recognition, Speech Processing (*i.e.*, Speech Synthesis, Speech Recognition), Pronunciation Generator, Morphological Analysis, Parts of Speech Tagging, Syntax, Grammar Checker, Text Categorization, Language Modeling and many more interesting research areas. In 2005, the Bangladesh Open Source Network (BdOSN: bdosn.org) was formed with local open source volunteers. BdOSN took Bangla in computing as one of its main issues. As a result, open source in Bangla has started to thrive (Islam, 2009). Though there are the various challenges leading to comprehensive use of Bangla in communication and information technology (ICT) such as –

- There is a lack of detailed morphological analysis of Bangla language.
- Need to build a larger and elaborate lexicon.
- Most of the current Bangla language computing tools are primarily based on Microsoft Windows operating system.
- There is a lack of coordination and integration among the various research groups working on the different areas of Bangla language processing.

In Bangladesh, we have few organizations, which are trying to work professionally such as Center for Research on Bangla Language Processing (CRBLP). The good news is that we have seen an explosion in the research papers being submitted to conferences that deal with Bangla language computing, especially in forums such as International Conference on Computer and Information Technology (ICCIT), an annual conference held in Bangladesh. The bad news is that there is still no significant synergy among the various research and development teams that are working in this field. But the lack of computational linguistics research on Bangla remains to be a significant challenge in Bangladesh. Researchers in neighboring India however have done noteworthy progress in this area and we hope to utilize much of that work and contribute to ongoing research.

The lack of consistent and targeted research funding remains to be a significant challenge in developing the skill-set that requires long-term development of Bangla language computational tools. The current environment is slowly changing to the intervention by the Government of Bangladesh through its Information and Computer Technology (ICT) initiative as well as externally sponsored research such as the one funded by Canada's IDRC in collaboration with NUCES in Pakistan. We hope to build a pool of local talents who are capable to conduct original research and development in this field and train the next generation of graduates from local universities.

However, many challenges remain in producing Bangla language computing tools and especially in integrating Bangla support into existing word-processing software, which is essential in making it usable to the larger body of users. Some of the challenges such as font design, input and output device support, linguistic analysis, lexicon and dictionary development are now actively being addressed.

8. Conclusion

The twenty-first century is the century of the total information revolution. The development of the tools for the automatic processing of the natural language spoken in a country or a whole group of countries is extremely important for the country to be competitive both in science and technology. To develop such applications, specialists in computer science need to have adequate tools to investigate language with a view of its automatic processing. One of such tools is a deep knowledge of both computational linguistics and general linguistic science. To expedite the basic research on Bangla language processing and Bangla language tools development, we need to address some important issues such as developing a local language dialect and indigenous language database and customized dialog system for easy access for all people, developing a good English-Bangla and Bangla-English machine translation system, developing a large and respective Bangla corpus and developing a

fully Unicode compatible Bangla operating system in Windows and Linux. In Bangladesh we see that some linguists take initiative personally with their own fashion but all of these are dying on the way due to the lack of sufficient financial support and materials. For the development and growing fascination about computational linguistic research in Bangladesh all computer science and linguistics departments both private and public universities and other non-government software organizations, should take initiatives.

References

- Ali, M. & Ali, M, M. (2002). Development of Machine Translation Dictionaries for Bangla Language. *Proceedings of 7th International Conference on Computer and Information Technology (ICIT)*, pp. 272-276.
- Grishman, R. (1994). *Computational Linguistics: An Introduction*. Cambridge: Cambridge University Press.
- Grishman, R., Macleod, C. & Meyers, A. (1994). Complex Syntax: Building a Computational Lexicon. In *Proceedings of the International Conference on Computational Linguistics, COLING-94*, Kyoto.
- Hirschman, L. & Sager, N. (1982). Automatic Information Formatting of a Medical Sublanguage. In *Sublanguage: Studies of Language in Restricted Semantic Domains*, R. Kitteridge and J. Lehrberger, eds.. Berlin: de Gruyter.
- H. Uszkoreit. (1997). *Language Technology (A First Overview)*. Survey of the State of the Art in Human Language Technology, Cambridge University Press and Giardini.
- Igor, A, Bolshakov. & A. Gelbukh. (2004). *Computational Linguistics: Models, Resources, Applications*. IPN – UNAM – Fondo de Cultura Económica, ISBN 970-36-0147-2, pp.187.
- Islam, M. S. (2009). Research on Bangla Language Processing in Bangladesh: Progress and Challenges. *8th International Language & Development Conference, 23-25 June 2009, Dhaka, Bangladesh*
- Lewis, M. Paul (ed.). (2009). *Ethnologue: Languages of the World*, Sixteenth Edition. Dallas, Tex.: SIL International. [Online] Available: http://www.ethnologue.com/ethno_docs/distribution.asp?by=country or http://en.wikipedia.org/wiki/Bengali_language.
- Montgomery, C. (1983). Distinguishing Fact from Opinion and Events from Meta-events *Proc. Conference Applied Natural Language Processing*, 1983, Santa Monica, CA, pp.55-61.
- O'Grady, William & Dobrovolsky, Michael. (1989). *Contemporary Linguistics: An Introduction*. St. Martin's Press, New York.
- Sager, N. (1978). Natural Language Information Formatting; the Automatic Conversion of Text to a Structured Data Basis. In *Advances in Computers*, 17, M.C. Yovits, ed., New York: Academic Press.
- Saha, G. K. (2005). The E2B Machine Translation: A New Approach to HLT. Ubiquity archive, *Association of Computing Machinery (ACM)*, 6 (32), New York.
- Slocum, J. (1984). Machine Translation: Its History, Current Status and Future Prospects. *Proc. Coling 84 (Tenth International Conference Computational Linguistics)*. Stamford, CA, pp.1-17.
- Slocum, J. (1985). A Survey of Machine Translation: Its History, Current Status and Future Prospects. *Computational Linguistics*, 11, 1 (Jan – Mar. 1985).
- Tucker, A. (1984). A Perspective on Machine Translation: Theory and Practice. *Comm. Assn. Computing Machinery*, 27, 4 (Apr. 1984), (pp. 322-329).
- Uddin, M. G., Ashraf, H., Kamal, A, H, M. & Ali, M, M. (2004). New Parameters for Bangla to English Statistical Machine Translation. *Proceedings of 3rd International Conference on Electrical & Computer Engineering (ICECE 2004)*, (pp. 545-548).

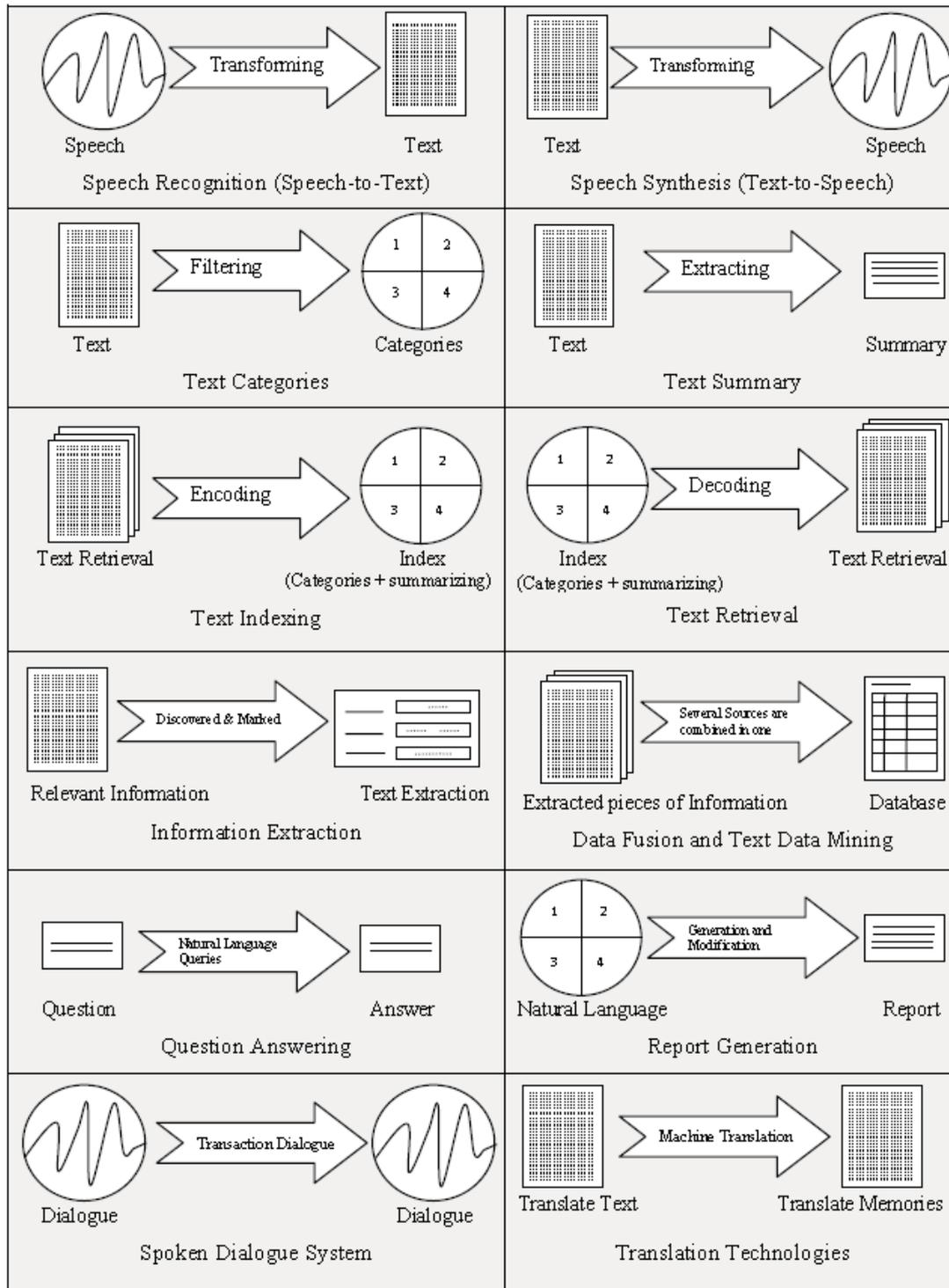


Figure 1. Functions of Language Technology

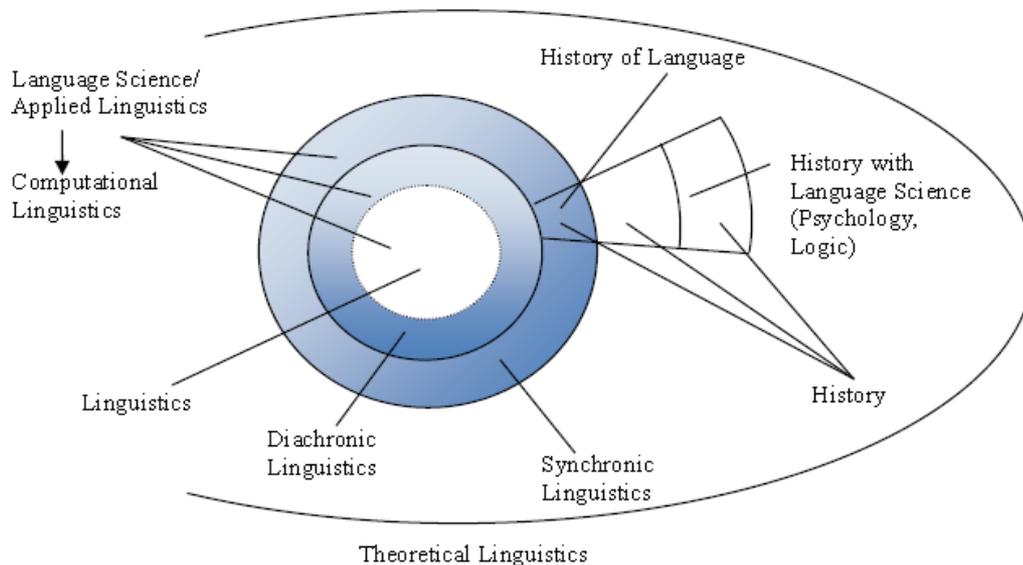


Figure 2. Position of Computational Linguistics in Linguistics Diversity

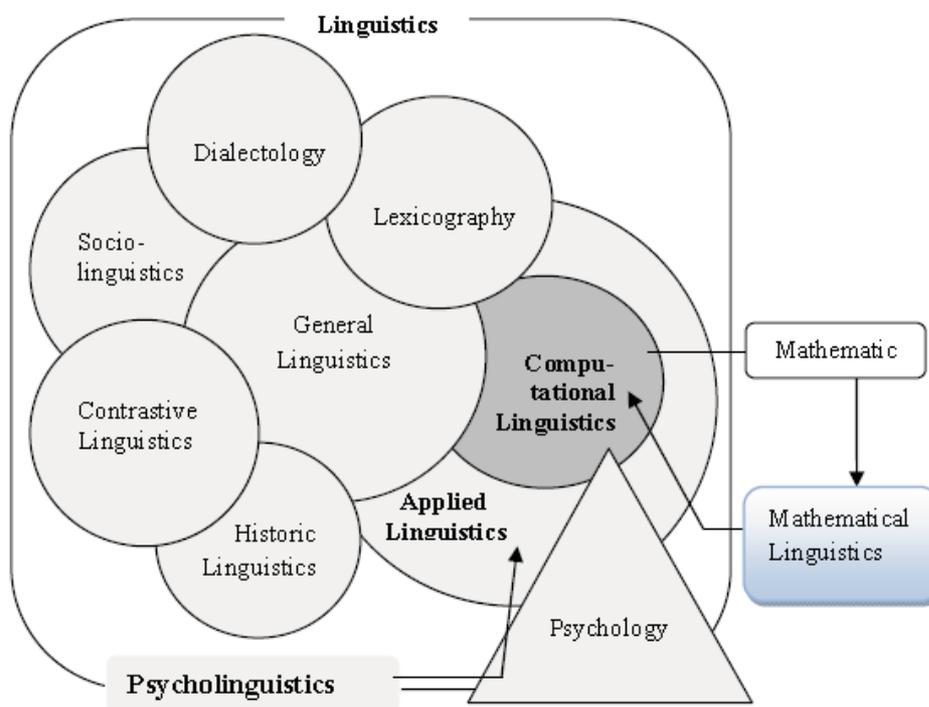


Figure 3. Structural Relationship between Linguistics Science & Computational Linguistics (Bolshakov & Gelbukh, 2004)

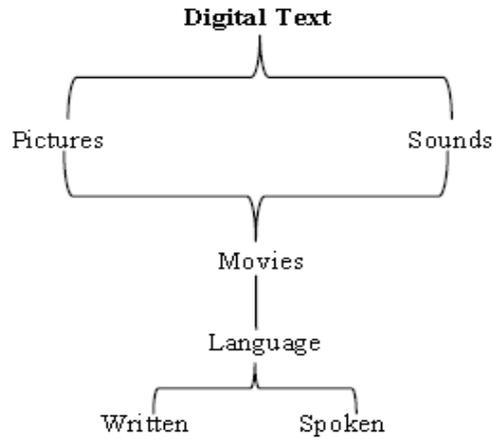


Figure 4. Text Combination

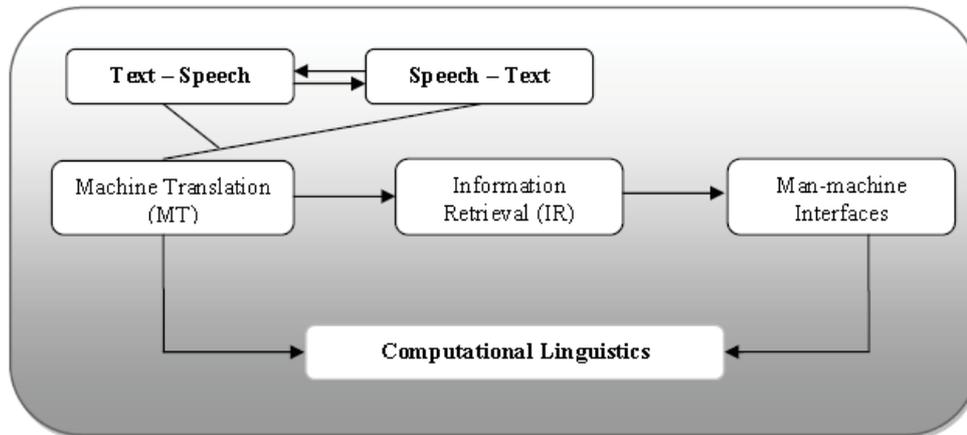


Figure 5. Central Development of Computational Linguistics

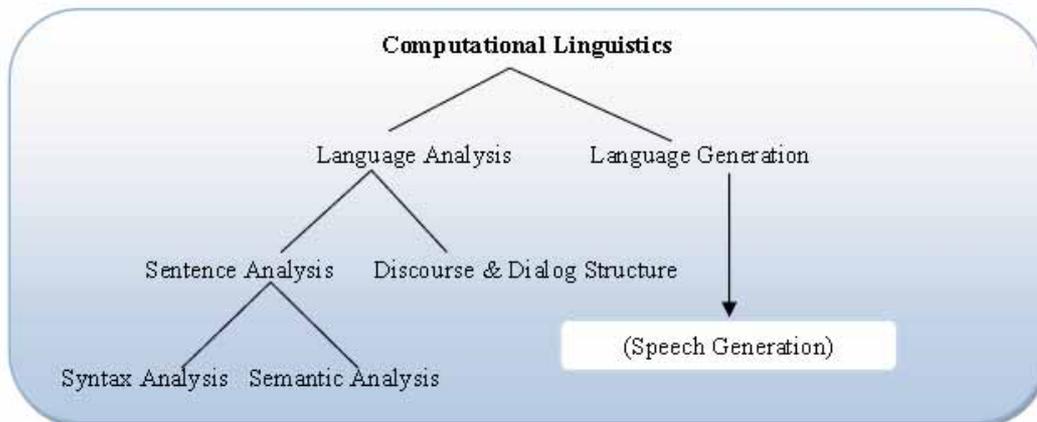


Figure 6. Structural Diagram of Computational Linguistics