К.А. Войтович
Беларусь, Брест, БрГУ имени А.С. Пушкина
Научный руководитель – Т.С. Троцюк

THE CONTRIBUTION OF COMPUTER SCIENCE TO LINGUISTICS AND TERMINOLOGY

The progress of computer science, particularly with the latest generations of personal computers and their penetration into the world of language, has brought about a change in the methods and tasks based on language processing. There are two potential roles for information technology (IT) in linguistics, just as in other areas: as a means of developing and testing models and as a means of gathering and analysing data. For example, one may use a computer to help to make a model of word formation properly specific, and also to gather and analyse some data on word forms. Thus, linguistics has the same types of use and benefit for computing as other academic areas, such as archaeology or economics.

The ties between linguistics and computer science have evolved gradually, and have given rise to various applications that can be classified according to the degree of complexity of the computational treatment applied. We can identify several stages in the development of this relationship.

In the first stage there were applications that were limited to using linguistic data as mere forms, without submitting them to any manipulation or analysis. Word processors (the most widespread of all microcomputer applications), spelling checkers, hyphenation programmes and communication programmes belong to this first stage.

The second stage includes machine-based linguistic tools designed for users working in language and communication: database managers, electronic dictionaries, systems to aid writers, translators or terminologists (computer-assisted systems for translation, writing, correction or learning).

In the third stage we find systems that manipulate data, either by analysing it or by converting it into data with other characteristics. Included here are analyzers, classifiers, programs for processing statistics, etc.

Finally, in the last stage we observe what are known as expert systems, which act with a certain amount of "intelligence" and to some degree attempt to replace human intervention. This stage includes automatic term identification and extraction, machine translation, systems for automated learning, automated indexing, text generation, etc. [1].

One more thing that needs to be mentioned is that data, or corpus, work is a natural arena for IT: computers can rapidly match, sort, count and so forth vast volumes of material.

Corpus use at the lowest, observational level appears to be referred to as one of the fastest growing areas of linguistics. It is illustrated, for example, by past uses of the Brown or Lancaster / Oslo-Bergen Corpora, and the use, especially by lexicographers, of the British National Corpus, Russian National Corpus and so on. It is hard to measure how valuable such browsing and observational use of simple word concordance and frequency data is, but the fact that serious publishers are willing to put money behind corpus construction suggests that corpora are seen useful and even essential.

Corpus work can be considered at three levels: observational, derivational, and validatory.

In the first, observational, case corpora can usefully display language phenomena, both recording and drawing attention to them. This was one of the earliest uses of IT for linguistic study, and remains important though as corpora get larger and it becomes harder to digest the concordance information.

The second, derivative level of corpus use is potentially much more interesting, but is also more challenging. It is aimed at a much more thorough analysis of data to derive patterns automatically: lexical collocations, subcategorization behaviour, terminological structure, even grammar induction. Such analysis presupposes first, some intuitive notion of the type of structure that may be present in the data and second, the actual algorithm for discovering model instances in the data.

The third level, theory validation, is where the two areas of IT utility for linguistics overlap. IT in principle offers great opportunities here, through making it possible to evaluate a theory of some linguistic phenomenon in a systematic, i.e. objective and comprehensive, way. IT is a stimulus to the development of formal theories. Moreover, the really important point about this work as a whole is that it has been closely tied to work on building systems for natural language processing (NLP) tasks, for instance translation or data extraction from text.

Computational linguistics has been developing strong interests in sublanguage studies in recent years. Computer-aided terminology can be located between computational linguistics (computer science applied to the processing of language data) and the industrialization of derived products (the language industries). The creation and development of computer-aided terminology can be accounted for by a series of factors related to computer science as well as to the needs of society itself. On the one hand, it has benefited from the progress and widespread use of text-based computer products using microcomputers. Today it is impossible to conceive of a language task being performed without computers. On the other hand, contemporary society, being dominated by the power of information and the need to communicate, places a great deal of importance on anything which facilitates communication. "Terminology, an essential element for specialized communication, is thus increasingly important as a means of transferring thought and technology" [1, p. 162].

We identify two different types of the influence of computer science on terminology: the introduction of computers in terminology has changed both the methodology of terminological work and the actual processing. Also the research in artificial intelligence has permitted terminologists to design expert systems that can perform some terminological tasks.

The effect of computer science on the methodology of terminology can be seen in the use of previously recorded electronic corpora and the exploitation of terminological and knowledge databases. Computer-aided text analysis and the possibility of processing large amounts of information have changed the bases of terminology compilation, as well as the degree of human intervention in the whole work process. The huge amount of data terminologists have available allows them to obtain well-founded information about terms and to have much other information at hand. On the whole, this provides a more solid foundation to the decisions terminologists have to make throughout the process.

There are five basic points in terminography at which computers can play a highly significant role for terminologists: selecting documentation, creating the corpus and extracting data, writing the entry, checking the information in the entry, ordering the terminological entries [1].

There are, nevertheless, still problems, of both a specific and a more general nature. The main weaknesses of the use of computers in terminological work are related to one or more of the following factors: lack of integrating computer resources in work methods; lack of compatibility among the resources themselves; the limited degree of computer processing available by each resource at present, as human intervention is constantly required; the inadequacy of such hardware items as text scanners; the limited number of existing corpora in different languages.

Despite these shortcomings, research in computerized terminology continues and increasingly attracts the attention of those investing in and working in computational linguistics and artificial intelligence. The development of computer applications has yet to produce the multipurpose, flexible products that can meet all the needs of language users and researchers.

1. Cabrè, M. T. Terminology: theory, methods, and applications / M. T. Cabrè ; translated by J. A. DeCesaris. – Philadelphia : John Benjamins Publishing, 1999. – 248 p.

В статье показывается влияние интенсивного развития информационных технологий на исследования в области лингвистики и терминологии. Представлены примеры использования информационных разработок в корпусной, коммуникативной лингвистике; показан вклад компьютерных технологий в методологию и практическую работу терминоведов.