Chapter 5

Automated translation systems: faults and constraints

Karolina Kuligowska, Paweł Kisielewicz, Aleksandra Rojek

Introduction

Automated translation, also known as machine translation, is based on automatically realized machine translation of text in one language (source language, SL) into text in another language (target language, TL). This field is also referred to as TTTL – Translate Text To Language.

Linguistic and philosophical ideas of creating a universal language and mechanical dictionaries date back to the seventeenth century. They remained a pure theorizing until the forties of the twentieth century, when technological improvements led to the first practical inventions. In 1949 Warren Weaver encouraged American scientists to build a computer-based translator. The first automatic translator, in a very basic form, was constructed in 1954 by researchers from Georgetown University in collaboration with IBM. The machine was able to translate at a time around sixty Russian sentences into English [Cheragui, 2012, p. 161]. This implementation started years of subsequent researches, concepts and discoveries in the field of machine translation. Nowadays, machine translation is present in everyday life and available at least for every Internet user [Hutchins, 1995, p. 431–445].

Automated translation system should be able to analyse all elements of a sentence in order to interpret its meaning and the context of used words. At the basic level, the system performs simple substitution of words. However this process cannot bring good results, because it is necessary to recognize whole phrases and their closest counterparts in the target language. Every natural language has its own grammatical structure and rules which have to be followed by machine translation systems. This requires extensive linguistic knowledge of grammar, syntax and semantics – not only in the source language, but also in the target one. For that reason, the biggest challenge is to create a translation module which could generate high-quality translations without the need of human intervention.

The aim of this paper is to present automated translation systems and to examine their drawbacks and limitations. The paper is organized as follows. Section 1 presents a brief review of machine translation approaches. Section 2 describes functioning of existing machine translation systems along with their architecture. Section 3 analyses faults and constraints of machine translation systems. Finally, the last section presents our conclusions.

5.1. Approaches to machine translation

Over the years of the development of machine translation, researchers have been adopting various approaches to this issue. The most general classification distinguishes three following approaches: rule-based approach, corpus-based approach and hybrid approach [Cheragui, 2012, p. 163–165; Langa, Wojak, 2011, p. 5; Tripathi, Sarkhel, 2010, p. 389–391].

5.1.1. Rule-based translation

Rule-based translation is based on a built-in set of linguistic rules, previously elaborated by linguists. This approach also includes gigantic bilingual dictionaries for each language pair. Rule-based translation system parses the source text and creates its temporary representation. Then, using a set of appropriate rules and transformations of grammatical structures, the temporary representation is reformulated into text in the target language. This process requires a comprehensive set of grammar and linguistic rules as well as extensive lexicons which contain morphological, syntactic and semantic information.

In this approach it is possible to achieve a good and very good quality of translation. The translation is coherent and predictable, even though it can be shorn of smoothness expected by readers. We have to be aware of the fact that the process of improving the quality of the translation has to be long and expensive. On the other hand, rule-based translation efficiency is high, even when realized on the standard hardware.

Rule-based translation constitutes the basis of following methods:

- direct translation approach,
- transfer-based approach,
- interlingual approach (i.e. translation using artificial intermediate language).

The clearest explanation of the complexity of the rule-based translation methods is presented on so-called Vauquois triangle illustrated in Figure 5.1.

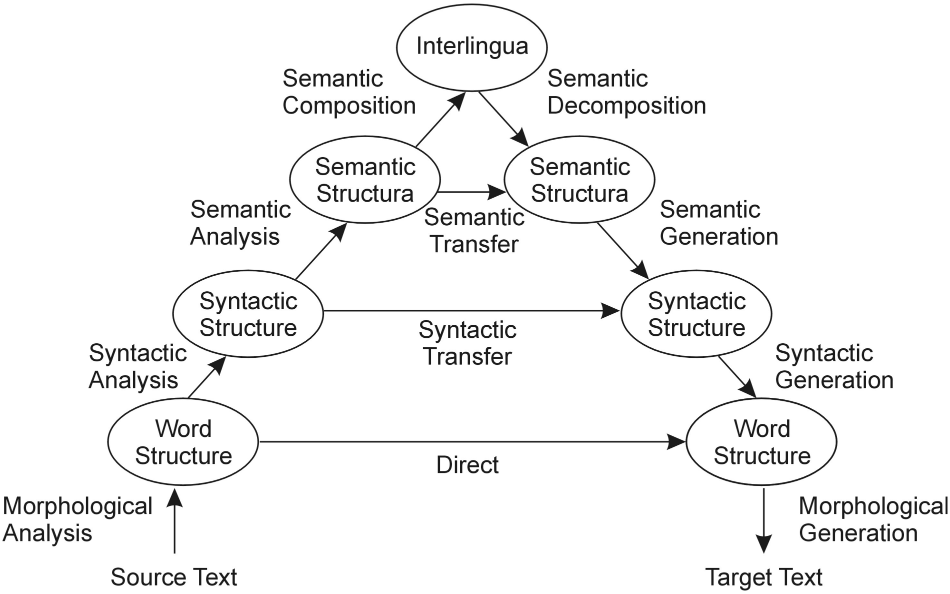

Figure 5.1. The Vauquois triangle

Source: [Dorr et al., 2005, p. 2].

The Vauquois triangle shows that the most basic method, which is direct translation, includes a low level of analysis and requires extensive knowledge about the structure of the word. When moving from the base to the apex of the triangle, one can observe an increase in the required level of analysis with a simultaneous decrease in the demand for knowledge about the structure of the word. Therefore selection of the method influences the level of depth of analysis on the one hand, and the extension of the knowledge and linguistic diversity on the other hand.

5.1.2. Corpus-based translation

Corpus-based translation is based mainly on existing multilingual corpora. They contain a minimum of 2 million words per specific field, and even more for colloquial language translations. In corpus-based translation approach it is possible to achieve a translation on a high level of quality, but sometimes the translation might be inconsistent and unpredictable. Unfortunately, most companies do not dispose of sufficiently large multilingual corpora, which are necessary in the process of building and training translation models. What is more, translation requires a significant amount of processing power. To achieve even average performance, expanded hardware configuration is necessary. However, if the company has adequately large multilingual corpora, process of quality improvement is fast and cost-effective.

Corpus-based translation is the basis of the following methods:

- statistical machine translation (SMT),
- example-based machine translation (EBMT).

Statistical machine translation uses statistical models that generate the most likely translation based on corpus, which is a large database of translated texts. In this method statistical correlation tables assign, on the basis of probability, words, sentences and phrases from the source language to their counterparts in the target language. Building statistical translation models is considered as a relatively fast process and does not require implementing a set of grammatical rules.

Example-based machine translation is a form of "translation by analogy" and it can be perceived as machine learning system which includes case-based reasoning. Examples are located in bilingual corpora, containing pairs of analogical sentences in the source and target language. These sentences simplify process of model training.

5.1.3. Hybrid translation

Hybrid translation combines strengths of previous approaches. Its aim is to achieve a very good level of translation quality and high efficiency on a given hardware (as in rule-based translation) while ensuring low investment costs (as in statistical translation).

5.2. Functioning and architecture of automated translation systems

5.2.1. Rule-based systems

In the direct translation system the translation process is based on the knowledge of the source language and knowledge about how to transform parts of analysed sentences in the source language to sequences of sentences in the target language. The architecture of this basic approach to automatic translation is illustrated in Figure 5.2.

On the other hand, in the transfer-based system translation requires extensive knowledge about the source and target languages, as well as about the connection between the analysed sentences in both languages. Therefore, the architecture of this system is also called linguistic knowledge architecture.

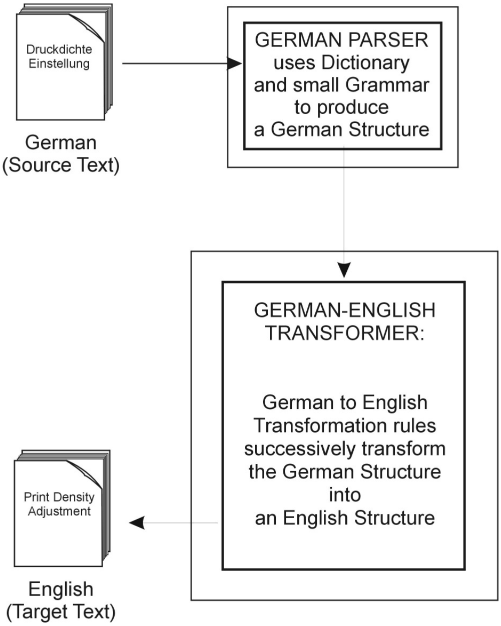
Druckdichte Einstellung
German (Source Text)
GERMAN PARSER uses Dictionary and small Grammar to produce a German Structure
GERMAN-ENGLISH TRANSFORMER:
German to English Transformation rules successively transform the German Structure into an English Structure
Print Density Adjustment
English (Target Text)

Figure 5.2. The architecture of a rule-based direct translation system
Source: [Arnold et al., 1994, p. 60].

The architecture of a transfer-based system is shown in Figure 5.3.

As it can be seen in the Figure 5.3, the architecture of a transfer-based system requires two components:

- analysis, which contains an impressive set of grammatical rules of the source language and the target language; these rules are used by parsers for the analysis of sentences in the source language and for transferring them into a symbolic representation,
- synthesis, which connects each representation of the sentence in the source language with a corresponding representation of the sentence in the target language. This representation is the basis for generating a translation in the target language.

The most complex rule-based system, i.e. interlingual system, represents a higher level of analysis than transfer-based approach. It uses so-called interlingua, an artificially created intermediate language. The architecture of an interlingual system is illustrated in Figure 5.4.

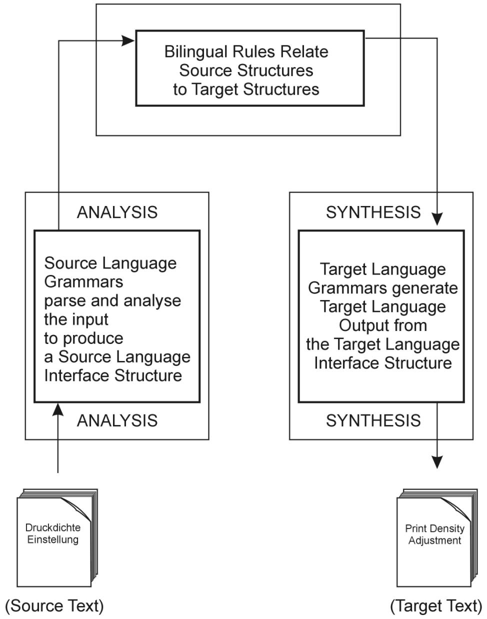
Bilingual Rules Relate Source Structures to Target Structures
ANALYSIS
Source Language Grammars parse and analyse the input to produce a Source Language Interface Structure
ANALYSIS
SYNTHESIS
Target Language Grammars generate Target Language Output from the Target Language Interface Structure
SYNTHESIS
Druckdichte Einstellung
(Source Text)
Print Density Adjustment
(Target Text)

Figure 5.3. The architecture of a transfer-based system

Source: [Arnold et al., 1994, p. 68].

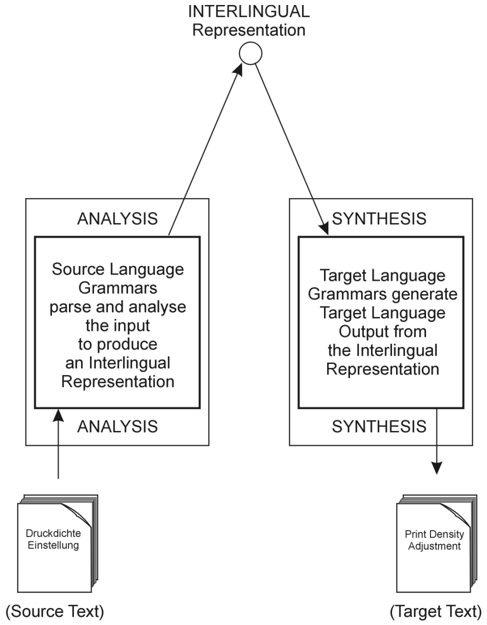
INTERLINGUAL Representation
ANALYSIS
Source Language Grammars parse and analyse the input to produce an Interlingual Representation
ANALYSIS
SYNTHESIS
Target Language Grammars generate Target Language Output from the Interlingual Representation
SYNTHESIS
Druckdichte Einstellung
(Source Text)
Print Density Adjustment
(Target Text)

Figure 5.4. The architecture of an interlingual system

Source: [Arnold et al., 1994, p. 79].

5.2.2. Corpus-based systems

Statistical machine translation (SMT) system makes a decision that is connected to probability. Among whole sentences in the target language, SMT syste-m must find the most likely translation of a given source sentence. The probability that the target sentence is the adequate translation of the source sentence is calculated on the basis of earlier learning of the model on small segments (sequences of words) of bilingual corpus of texts. Thus, in this approach whole translation constitutes a sum of shorter fragments translation.

Figure 5.5 illustrates the basic architecture of the SMT system. SMT translation involves two main stages: 1) training, during which the system is taught from available translation examples, and 2) testing, during which the new sentences are translated.

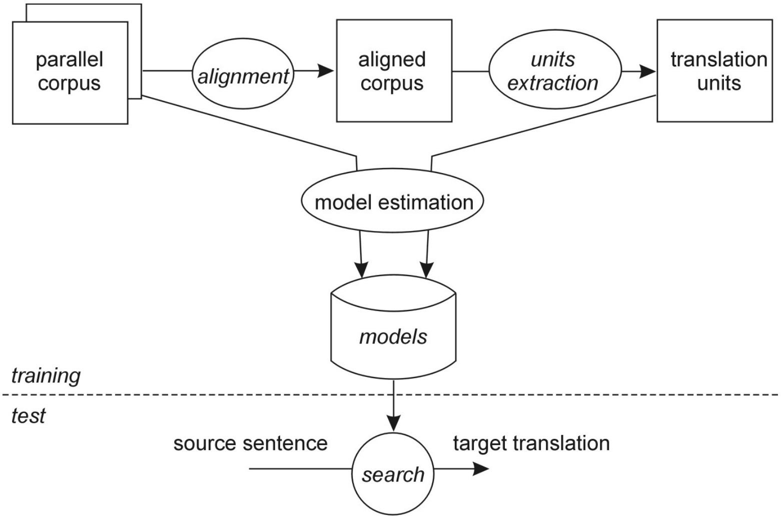

Figure 5.5. The architecture of statistical machine translation system

Source: [Crego Clemente, 2008, p. 5].

Training stage starts with sentence-to-sentence parallel corpus alignment, and continues with determining pairs of counterparts generated automatically by word-to-word alignment. This process is also called as "text binding" [Lewandowska-Tomaszczyk, 2005, p. 43]. Then, translation units (phrases) are automatically extracted from a parallel training corpus. They are used in the testing stage while generating new sentences. Finally, the last stage is phrase scoring. In this step, the translation probabilities are computed and scored for all phrase pairs [San-Segundo et al., 2013, p. 66]. While searching for a sentence with the highest translation probability, there are used several models responsible for adequacy and smoothness of translation.

Example-based machine translation (EBMT) system is based on intuitive assumption that people use already existing examples of translations in order to translate new input. If such system has to function properly, it must include bilingual corpus of parallel examples (their other name is "bitext" or example-base) to translate each part of a sentence. EBMT is based on previous translations in order to generate further translations. This process is broken down into three stages:

- matching,
- alignment,
- recombination.

The matching module finds an example or a set of examples from a parallel corpus, that matches best to the sequence of words in source language. The alignment module identifies equivalents within the string of "source-target" words from examples extracted previously during the matching stage. Recombination generates the final translation by putting together essential parts of the translation in the target language. Figure 5.6 illustrates the architecture of the example-based machine translation system.

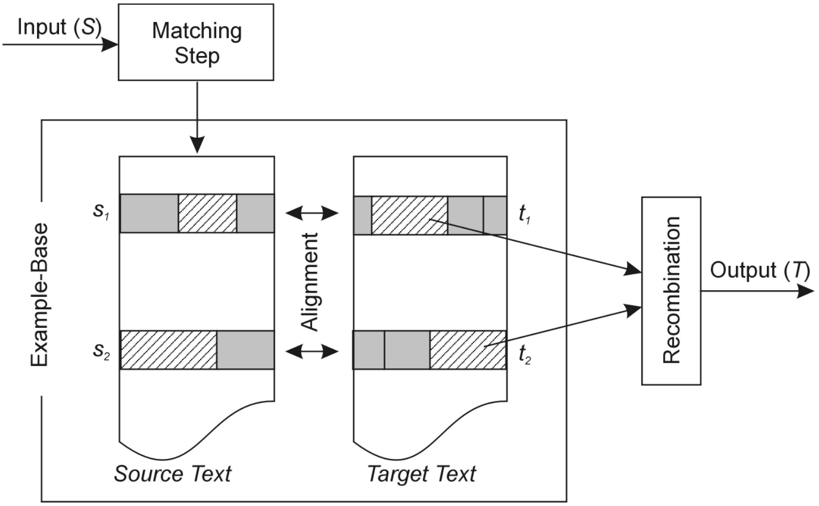

Figure 5.6. The architecture of the example-based machine translation system
Source: [Dandapat, 2012, p. 12].

5.3. Faults and constraints of automated translation systems

Automated translation systems make a lot of mistakes that almost never occur in case of human translations. Professional translator is more aware of the context and other important aspects of the translation. All of these limitations related to machine translation differ depending on the translated language pair. When two translated languages are profoundly dissimilar, then these errors are critical.

5.3.1. Ambiguity

One of the problems in machine translation is ambiguity which usually refers to vocabulary and language structure. What is more, those two types of ambiguity have the most impact on the quality of the translation. Ambiguity related to vocabulary occurs when words have multiple meanings. Ambiguity of the language structure results from the fact that we can interpret the same sentence in multiple ways.

5.3.2. Accuracy

Accuracy in machine translation is not always on the same level. Most systems translate word for word without understanding translated information. If we do not take into account the meaning of the text, the most common result is a general outline of the translation. Such translation has to be corrected by human.

5.3.3. Context

Automated translation systems cannot use previous experience in a way that human translators do it. In languages such as English, one word can have hundreds of different meanings depending on the context. Therefore, in order to properly translate the text, a system should take into account the intersentential context. As a result we should obtain a coherent multi-sentence text in target language. Unfortunately, machine translation has not yet reached the level that would allow to understand the whole intersentential syntactic context.

5.3.4. Training corpuses and machine learning

Performance of machine translation depends on training phase, therefore it might suffer from the problem of data sparseness. In case of small amount of training set, there is a problem of data prediction. This requires wide coverage of full lexicon of source and target language. Problem also occurs when the training data is from one domain, and system is applied to operate in some other field. Although parallel corpora are becoming increasingly available for various language pairs, the size of such corpora for the most language pairs are limited [Sugandhi et al., 2011, p. 3].

In order to achieve machine translation quality close to the human translation level, translation models must be able to recognize complex syntax and semantic representations and their equivalents in various languages. The key task of model learning in this case is to identify correspondence between structures in two languages and to model these connections statistically.

More difficult thing is to teach models using parallel data, where the syntax and semantic structures are available only for one of the two languages. The model learning task in this case is to project structures from one language to the their corresponding structures in the second language using word-to-word correspondence. The problem lies in obtaining needed training resources. The development of annotated training corpora is a necessary step without which researches on the development of such machine learning system cannot even begin [Sugandhi et al., 2011, p. 5].

5.3.5. Language constraints

There are four most important language constraints which occur during machine translating process [Sugandhi et al., 2011, p. 4–5]:

1) cultural differences,
2) changes in linguistic theory,
3) morphological complexity,
4) researches focused only on English.

Cultural differences

Cultural differences are a problem in current machine translation systems. The bigger difference between the source and target culture, the more problems arise while translating. Some words associated with one culture do not have their equivalent words in other languages. For example, the Indian word "sari" (traditional dress of Indian women) often has no equivalent word in other languages [Sugandhi et al., 2011, p. 4]. Here occurs a problem of untranslatability of individual words, with which often cannot cope even professional translators.

Another language problem, in the context of cultural differences, constitute unknown words. Unknown words are the source language words which are not included in the training data, and thus have no equivalent in the target language. The current machine translation systems usually omit such words, or leave them in their original form. This is justified in the case of names (assuming you do not have to make transliteration), but neither the omission of the word, nor leaving it in its original form, is a satisfactory solution. It is well known that unknown words issue significantly influences the quality of translation. This problem can be particularly severe when the available bilingual data is very small. First of all, it is difficult to find the meaning of these words in the target language. What is more, the unknown word can negatively affect the lexical selection and reordering the words around it. The conventional solution to the problem of unknown words is to find their equivalents in the target language with additional resources, such as multilingual data, web data or linguistic resources such as WordNet. However, most of these methods can cope only with certain types of unknown words, such us named entities, abbreviations, compounds or morphological variants. Therefore, problems connected with unknown words still remain unsolved. Moreover, the translation of such words with which the system copes, might not help in the lexical selection and reordering of the surrounding words because translation is obtained for other re-sources than the original bilingual training data [Zhang et al., 2012, p. 176–178].

Another constraint which depends on culture is translation of idioms. Idioms are defined as a multi-words expressions with constant sense (often metaphorical), meaning of which cannot be fully understood from the individual meanings of its elements. One of the issues that concerns precisely this kind of expressions is their ambiguity, so thus they can be interpreted literally and metaphorically. Some part of idioms can be translated word for word, provided that a similar expression exists in the target language. However, in cases when there is no similar idiom in target language, such translation is not possible. Very often this type of expressions are culturally limited, which means that they can exist only within one country or even a small region. Therefore, it is very difficult to transfer them into totally different cultural context. One method for translating these idioms is to find idioms in the target language with a similar meaning and form, or similar meaning, but different form. It is also possible to use a paraphrase. If idioms have no close counterparts, system can simply omit them [Gaule, Josan, 2012, p. 51]. Unfortunately, in many cases systems still cannot cope with this kind of expressions. Idioms require specific, separate rules, obviously in addition to the standard rules that apply to “ordinary” words and other linguistic structures.

Changes in linguistic theory

The development of machine translation, considered from the point of view of system performance, has to go hand in hand with the development of linguistic theory. However, this is difficult to achieve in practice, since modification of the knowledge base is not easy, especially when it comes to the colloquialisms. This results in a large communication gap between theoretical linguistics and practical research in machine translation. We also must take into consideration acronyms and official words which we use in translation, especially in multilingual translations. Acronyms are difficult to translate due to the fact that different letters are used in various languages, often in a changed order. This can be avoided in the future if they will be replaced by acronyms accepted in many languages [Sugandhi et al., 2011, p. 4].

Morphological complexity

Each language has a different level of morphological complexity. When it comes to English, the morphology is quite simple. Therefore, recent research in machine translation has paid only limited attention to issues of effectively handling complex morphology. All methods of machine translation tend to retain the structural characteristics of the source language, despite the fact that they should be more oriented toward the target language. In a multilingual machine translation various methods of translations for different target languages are implemented. To change this, researchers have been developing new translation models that effectively cope with complex morphology. The issue of data sparseness should be also taken into account [Sugandhi et al., 2011, p. 4–5].

In the source literature there are also mentioned several open and long-term challenges that need to be solved in the near future of automated translation [Lopez, Post, 2013, p. 2]:

- translation of sparse language pairs,
- translation across different domains,
- translation of informal text,
- translation into morphologically rich languages.

Researches focused only on English

USA is the heart of development of computer technology. However, this is one of the most homogeneous societies in the world in terms of language. For this reason, most of the linguistic theory of machine translation is based on phenomena observed in English. In addition, statistical machine translation is focused on a small number of language pairs for which huge amounts of sentence-aligned parallel texts have become available. Theories such as Lexical Functional Grammar or Generalized Phrase-Structure Grammar and their various derivatives were intended to cover as large range of languages as possible, not only within one specific language, but also for different types of languages.

When comparing various languages to English, they differ in a writing system, grammatical structure, and in a way of expressing similar meanings and intentions. Many world languages use some variant of the Latin alphabet, including particular special characters, or a totally different writing system than English. Additionally, languages such as Arabic, Persian and Hebrew are written from right to left, while Japanese and Chinese may be written from top to bottom [Sugandhi et al., 2011, p. 4–5]. All these issues need innovative computing solutions.

A particular challenge in machine translation are less popular European languages, which are characterized by rich morphology and language structure. For economic reasons, current commercial activities focus exclusively on the most widely spoken languages. For example, despite the fact that Google Translate covers more than 70 languages, the quality of translation of e.g. Baltic languages is much worse than English, French, Spanish. Languages that are not widely used, need special attention and detailed researches.

Conclusion

Although machine translation is a rapidly developing technology, there are still some limitations in the current automated translation systems. They mainly relate to variability of semantic meanings, which is conditioned by historical, cultural and civilizational factors. Automated translation problems also arise from the syntax differences between source and target language. In addition, machine translation systems often cannot cope with homonyms, synonyms and metaphors. There are also other issues to encounter, such as: recognition of the context of translated sentence, cultural differences between pairs of translated languages, changes in linguistic theories and morphological complexity of many natural languages. Unfortunately, extensive development of machine translation is not facilitated by researches strongly focused on English, which is the twenty-first century Latin.

References

1. Arnold D.J., Balkan L., Lee Humphreys R., Meijer S., Sadler L. (eds.) (1994), *Machine Translation: an Introductory Guide*, Blackwells-NCC, London.
2. Cheragui M.A. (2012), *Theoretical Overview of Machine translation*, in: Malki M., Benbernou S., Benslimane S.M., Lehireche A. (eds.), *Proceedings of the 4th International Conference on Web and Information Technologies (ICWIT)*, "CEUR Workshop Proceedings", Vol. 867.
3. Crego Clemente J.M. (2008), *Architecture and modeling for N-gram-based statistical machine translation*, doctoral dissertation, Department of Signal Theory and Communications, BarcelonaTech (UPC), Barcelona.
4. Dandapat S. (2012), *Mitigating the Problems of SMT using EBMT*, doctoral dissertation, Dublin City University School of Computing.

5. Dorr B.J., Hovy E.H., Levin L.S. (2005), *Machine Translation: Interlingual Methods*, in: Brown K. (ed.), *Encyclopedia of Language and Linguistics*, 2nd ed., Elsevier.
6. Gaule M., Josan G.S. (2012), *Machine Translation of Idioms from English to Hindi*, "International Journal of Computational Engineering Research", Vol. 2, Issue 6.
7. Hutchins W.J. (1995), *Machine translation: a brief history*, in: Koerner E.F.K., Asher R.E. (eds.), *Concise History of the Language Sciences: From the Sumerians to the Cognitivists*, Pergamon Press, Oxford.
8. Langa N., Wojak A. (2011), *Ewaluacja systemów tłumaczenia automatycznego*, master's thesis, Wydział Matematyki i Informatyki Uniwersytetu im. Adama Mickiewicza w Poznaniu, Poznań.
9. Lewandowska-Tomaszczyk B. (ed.) (2005), *Podstawy językoznawstwa korpusowego*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.
10. Lopez A., Post M. (2013), *Beyond Bitext: Five open problems in machine translation*, Twenty Years of Bitext.
11. San-Segundo R., Montero J.M., Giurgiu M., Muresan I., King S. (2013), *Multilingual Number Transcription for Text-to-Speech Conversion*, 8th ISCA Workshop on Speech Synthesis (SSW-8), ISCA, Barcelona.
12. Sugandhi R., Charhate S., Dani A., Kawade A. (2011), *Addressing Challenges in Multilingual Machine Translation*, "International Journal of Scientific & Engineering Reaserch", Vol. 2, Issue 6.
13. Tripathi S., Sarkhel J.K. (2010), *Approaches to machine translation*, „Annuals of Library and Information Studies", Vol. 57.
14. Zhang J., Zhai F., Zong C. (2012), *Handling Unknown Words in Statistical Machine Translation form a New Perspective*, in: Zhou M., Zhou G., Zhao D., Liu Q., Zou L. (eds.), "Natural Language Processing and Chinese Computing (NLPCC 2012), Communications in Computer and Information Science", Vol. 333, Springer-Verlag Berlin Heidelberg.