TEXT CORPORA AND ITS ROLE IN DEVELOPMENT OF THE LINGUISTIC TECHNOLOGIES FOR THE AZERBAIJANI LANGUAGE

Sevinj Mammadova[1], Gulnar Azimova[2], Abulfat Fatullayev[3]

[1,2]National E-Governance Network Initiative Project, Baku, Azerbaijan
[3]Cybernetics Institute of ANAS, Baku, Azerbaijan
[1]*sevinc@dilmanc.az*, [2]*gulnar@dilmanc.az*, [3]*fabo@box.az*

Abstract. In this paper the role and importance of the text corpora for the development of the linguistic technologies are described. We have introduced creation process of the first version of Azerbaijani monolingual and bilingual text corpora.

1. Introduction

Development of Natural language Processing (NLP) systems is a foreground field of the Computer Sciences. The desire to use the computers in all fields of the human activity and to interact with the computers by more natural ways such as speech demands the development of different linguistic computer technologies – machine translation (MT), automatic speech recognition (ASR), text-to-speech (TTS), speech understanding-generation and other systems.

The history of development of formal linguistic technologies for Azerbaijani can be divided into two parts: before and after 2003. The researches conducted till 2003 were mostly theoretical and no software was created during this period. M.Mahmudov's works (Mahmudov, 2002) are considered to be the most important among these research works [1]. Beginning from 2003 the development of applied linguistic technologies for Azerbaijani has already been started. The research works on the development of applied linguistic technologies for Azerbaijani are being implemented within the frame of Azerbaijan Ministry of Communications and Information Technologies and UNDP-Azerbaijan joint project (Dilmanc project). As a result of research works Dilmanc Machine Translation system and Dilmanc ASR and TTS system were developed [2-5]. Currently, Dilmanc Telephone Translator system, Text Corpora for Azerbaijani language are being researched and developed. In this research works we use both rule-based and statistic approaches respectively.

Despite indisputable success in the theory and practice, existing program products on linguistic technologies are far from the perfection. Since the nature of human brain function is explored to limited extent only, the task to create the software that imitates the functions of human brain is very complicated.

For this reason in the case of MT we can operate only with input and output data without understanding the transformation process (Fig. 1).

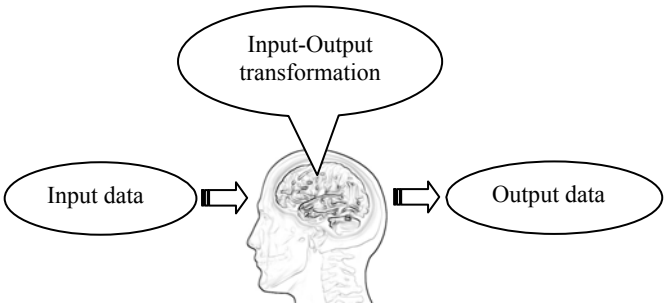

Fig. 1.

Because we do not know the nature of transformation for modeling this process we can use input and output data as it is mentioned above, i.e. what produces the definite input data. By involving the statistics it is possible to solve this problem with the definite accuracy.

According to linguistic technologies for obtaining statistics it is necessary to create the large text corpus. In linguistics, text corpus is a large and structured set of texts (now usually electronically stored and processed). They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within specific spheres.

2. Types of Text Corpora

There are two types of text corpus: Monolingual corpus is the corpora that contains texts in a single language and multilingual corpus which is the text data in multiple languages. Multilingual corpora that have been specially formatted for side-by-side comparison are called aligned parallel corpora [6].

The corpora are often subjected to a process known as annotation in order to make it more useful for doing linguistic research [6]. Part-of-speech tagging or POS-tagging is an example of annotating a corpus. During this annotation information about each word's part of speech (verb, noun, adjective, etc.) is added to the corpus in the form of tags. Another example is indicating the lemma (base) form of each word [7].

When applying analysis some corpora have further structured levels. Especially, a number of smaller corpora may be fully parsed and these corpora are usually called Treebank or Parsed Corpora. Treebank are usually used in computational linguistics for training or testing parsers or in corpus linguistics for studying syntactic phenomena. The Treebank is usually smaller and it can contain around 1 to 3 million words. Linguistic structured analysis includes different levels such as annotations for morphology, semantics and pragmatics. In corpus linguistics, corpora are the main knowledge base [8].

Corpus Linguistics has generated a number of research methods, attempting to trace a path from data to theory. Annotation, Abstraction and Analysis were presented as the "3A perspective" in 2001 by Wallis and Nelson [9]. The structure of 3A perspective components is the following:

- Annotation is the application of a scheme to texts. Structural markup, part-of-speech tagging, parsing, and numerous other representations may be the elements of Annotations.
- Abstraction consists of the translation (mapping) of terms in the scheme to terms in a theoretically motivated model or dataset. Abstraction usually includes linguist-directed search but sometimes may include rule-learning for parsers.
- The components of Analysis are statistically probing, manipulating and generalizing from the dataset. Statistical evaluations, optimization of rule-bases or knowledge discovery methods may also be included in Analysis.

The part-of-speech-tagged (POS-tagged) is the most common lexical corpora today. But even corpus linguists who work with 'unannotated plain text' inevitably apply some method to isolate terms that they are interested in from surrounding words. In this case they combine annotation and abstraction in a lexical search. The positive aspect of publishing an annotated corpus is that other users can then make editions on the corpus. Linguists with other interests and differing perspectives than the originators can exploit this work. By sharing data, corpus linguists can treat the corpus as a locus of linguistic debate, rather than as an exhaustive fount of knowledge.

2. Examples for Text Corpora

We would like to review some famous and the most used text corpora below:

The American National Corpus (ANC) is a text corpus of American English. It currently contains 22 million words written and spoken data produced since 1990. The ANC includes a range of genres comparable to the British National Corpus and is annotated for part of speech and lemma, shallow parse, and named entities (www.americannationalcorpus.org) [10].

The British National Corpus (BNC) by using a wide range of sources was able to collect a 100-million-word text corpus of samples of written and spoken English. It was compiled as a

general corpus in the field of corpus linguistics. BNC intended this corpus to be a representative sample of spoken and written British English of the late twentieth century from a wide variety of genres (www.corpus.byu.edu/bnc) [11].

The Russian National Corpus (Национальный корпус русского языка) was shared online since April 29, 2004 and is available at the moment. The initiator of this project is the Institute of Russian language, Russian Academy of Sciences. It currently contains about 150 million word forms that are automatically lemmatized and POS-/grammeme-tagged, i. e. all the possible morphological analyses for each orthographic form are ascribed to it. Lemmata, POS, grammatical items and their combinations are easily searched. Moreover, 6 million word forms are in the subcorpus with manually resolved homonymy (www.ruscorpora.ru) [12].

3. Azerbaijani text corpus

For the creation of the monolingual corpus majority of Azerbaijani-language Internet sites and other electronic resources in Azerbaijani are used. On the base of these works we could gather the Azerbaijani texts consisting of about 300 million word-forms. All files are downloaded in the *php*, *asp* and *js* formats and transformed into *html* format. After this process the files are transformed into *rtf* format by using HTML-TO-RTF converter and html tags are removed. Finally, all files are transformed into *txt* format by using RTF-TO-TXT converter and edited shallowly. In editing process the special software developed by programmers of the Dilmanc R&D group also are used.

Presently, this corpus is used for building the language model for Azerbaijani ASR and MT systems.

By the same way texts are collected for the formation of Azerbaijani-English bilingual corpus. After downloading the files and their English equivalents are aligned. The corpus is consisting of more than 1 million sentences. Currently, members of Dilmanc R&D group work on the alignment the sentences of the bilingual corpus. More than 30000 sentences are already aligned and this work is continued (Table 1).

Table 1. **Fragment of the aligned corpus**

Azerbaijan is one of the oldest spots of civilization, a country with a rich and ancient history.	Azərbaycan sivilizasiyanın ən qədim mərkəzlərindən biri olmaqla zəngin və qədim tarixə malikdir.
Dissemination of information on the Council of Europe, its activities and standards;	Avropa Şurası, onun fəaliyyəti və standartları haqqında məlumatları yaymaq;
Applications of citizens to state authorities with proposals, statements and complaints shall be the main aspect of human rights practices and protection.	Vətəndaşların dövlət orqanlarına təkliflər, ərizə və şikayətlərlə müraciət etməsi insan hüquqlarının həyata keçirilməsində və qorunmasında mühüm vasitədir.
From that moment on, we are in a race against time.	Bu andan etibarən biz zamanla mübarizə aparacağıq.
The most ancient inscriptions related to the Caspian Sea were encountered upon an Assyrian ceramic pot, mentioned as the Southern Sea.	Xəzər dənizi haqqında ən qədim məlumatlara Assuriya gil qablarının üzərindəki yazılarda rast gəlinir və o Cənub dənizi adlanır.
…	

4. Conclusion and future works

As mentioned above monolingual and bilingual corpora have a great importance from the view of the development of the linguistic technologies. Created corpora should be used while developing the applied linguistic technologies for Azerbaijani in general. Particularly, as a result of providing research and development works on the development of the corpora the statistic component of the hybrid MT system will be improved and the recognition accuracy of the Dilmanc ASR system will be increased.

Currently, the works on the alignment, POS-tagging and syntactic tagging of Text Corpora for Azerbaijani language are being continued.

References

[1] Mahmudov M. 2002. Metnlerin formal tehlili sistemi. Elm, Baku.
[2] Fatullayev R, Abbasov A, Fatullayev A. 2008. Dilmanc is the 1st MT system for Azerbaijani. In: Proc. of SLTC-08, Stockholm, Sweden, pp. 63-64.
[3] Fatullayev R, Mammadova S, Fatullayev A. Translating Composite Sentences in Azerbaijani-English MT System. ISMTCL - International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages, and their application to emergencies and safety critical domains, July 1-3, 2009, pp. 138-142.
[4] Fatullayev R, Abbasov A, Fatullayev A. 2008. Peculiarities of the development of the dictionary for the MT System from Azerbaijani. In: Proc. of EAMT-08, Hamburg, Germany, pp. 35-40.
[5] Abbasov A, Fatullayev A. 2007. The use of syntactic and semantic valences of the verb for for-mal delimitation of verb word phrases. In: Proc. of L&TC'07, Poznan, Poland, pp. 468-472.
[6] Sinclair, J. 'The automatic analysis of corpora', in Svartvik, J. (ed.) Directions in Corpus Linguistics (Proceedings of Nobel Symposium 82). Berlin: Mouton de Gruyter. 1992.
[7] *Garside, R., Leech, G. N., and McEnery, T. 1997.* Corpus annotation: linguistic information from computer text corpora*. London: Longman.*
[8] Baker, J. P. 1997. Consistency and accuracy in correcting automatically tagged data. In Corpus annotation: Linguistic information from computer text corpora*, eds. Roger Garside,* G. Leech and A. McEnery, 243-250. London: Longman.
[9] Wallis, S. and Nelson G. 'Knowledge discovery in grammatically analysed corpora'. Data Mining and Knowledge Discovery, 5: 307-340. 2001.
[10] Fillmore, Charles, Nancy Ide, Dan Jurafsky, and Catherine Macleod. 1998. An American National Corpus: A Proposal. Proceedings of the First Annual Conference on Language Resources and Evaluation. Paris: European Language Resources Association, 965--969**.**
[11] Aston, Guy and Lou Burnard 1998. The BNC Handbook: Exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press.
[12] Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009.