Culture & Creativity

EU-Eastern Partnership Programme

DEVELOPING AN ONLINE DICTIONARY AS A NATIONAL TREASURY

A policy brief for Culture & Creativity
EU-Eastern Partnership Programme

07/12/2017

_____

Gunel Rzayeva

Executive Summary

The present Policy Brief discusses development of Azerbaijani language through creation of online dictionaries.

Every two weeks the world loses one language in average. The Azerbaijani language will only survive in our globalised information and knowledge society if the publishing industry can produce high quality and high quantities materials published both in hard and soft copies. The Government of Azerbaijan has always realized the vital role of language for nation building and informational exchange between nations and generations and therefore, always paid special attention to this topic. In 2013 President of Azerbaijan Republic approved the «State Programme on Use of Azerbaijani language in globalization environment in accordance with requirements of the era and Development of linguistics in the country». There is wide use of ICT in research and promotion of Azerbaijani language and specifically preparation of new and improved electronic dictionaries of Azerbaijani language among the five priorities of the Programme.

Analysis of current situation with dictionaries in AZ.net showcases 11 key areas that the publishing industry in Azerbaijan needs to focus on in order to improve the current offer of dictionaries.

It is recommended considering development of an open edited dictionary web resource within the broader context of establishing a national movement for ensuring the purity and beauty of the Azerbaijani language, called *Our Language is a Treasure* (*Dilimiz Sərvətimizdir*). The general idea that would underline the movement would be: *A Small Contribution Can Make a Big Difference.*

Introduction

Every two weeks the world loses one language in average. There are over 7,000 languages on the Earth. But only 80 of them are spoken by 80% of the world population. The more communications develop, the less languages survive. Currently, over 400 languages are considered endangered. About a half of the existing languages will no longer be spoken already by mid-XXI century. 1

UNESCO published «The Red Book on Endangered Languages» and an Atlas of the World's Languages in Danger, which are also available online. In order to assess language sustainability these reports took the following factors into account:

- Stable orthography in the established writing system;
- Reference literature (grammar books, dictionaries, classical literature);
- Mass dissemination materials (printed media, audio records, movies, songs and music);
- Science and education literature (scientific publications, research projects, lesson-books, manuals);
- Various carriers of everyday information (posters, advertisements, correspondence, reference information, guidelines etc.);

In other words, one of the main language survival factors is the availability of an adequate volume of quality information resources allowing the uninterrupted transmission of language between generations (i.e. ensuring language is spoken by all generations).

Migration and globalisation are the main reasons for language extinction. The era of globalization has affected virtually all areas of our life including such an important component as language. Globalization's effect on language is that many languages become extinct as their speakers come

---

into contact with a stronger language environment. This is particularly true about the languages of small ethnic groups.

Everyone in Azerbaijan has the right to freely choose the language in which to get information, whether that is to review the news or read classics. Azerbaijani resources need to be of the same quality as others widely used in Azerbaijan including Russian, English and Turkish in order to ensure that Azerbaijani-language materials are preferred. The Azerbaijani language will only survive in our globalised information and knowledge society if the publishing industry can produce high quality and high quantities materials published both in hard and soft copies.

Current Situation

Review of main electronic English-Azerbaijani dictionaries

Let us look at the most popular electronic dictionaries in Az.Net focusing on English-Azeri-English translation dictionaries. A more detailed comparison table of all of these dictionaries can be found in the Annex.

 Polyglot was the first electronic dictionary to be produced in Azerbaijan in 2004. It is available in both desktop and online versions. Due to the widespread use of the desktop version, the online dictionary has not been active for a long time. Along with English, Polyglot includes French, German and Russian dictionaries. It is easily accessible (downloadable from the website and its size is only 20Mb) and therefore very popular. The main weakness is that it is based on an outdated paper dictionary and has not been updated since its creation 13 years ago. For example, if you search for the word *computer* you will get the translation e*lectronic computing machine* in English-Azerbaijani and not get any English equivalent of the Azerbaijani word "*kompüter*". The bibliography includes hard copy dictionaries with a total number of 68,500 entries. However, these dictionaries will overlap and the actual total number of words will be much lower.

 The ambitious "Dilmanc" project was launched in 2006 to develop translation software between Azerbaijani and English, Russian and Turkish within the National Governance Initiative implemented by the Ministry of Communications and Information Technologies in partnership with UNDP. The project developed:

- A software translation system (Azeri-English-Azeri, Azeri-Russian-

Azeri and Turkish-Azeri) with the capacity to translate texts of up to 200 words immediately.

- A speech recognition system for Azerbaijani
- A text-to-speech system for Azerbaijani
- And work on improving a phone-translator application is in progress

The Dilmanc website also has a dictionary section that is based on machine translation rather than human-written dictionaries and therefore does not guarantee correct translation. A significant advantage of the dictionary is its ability to recognize word forms of Azerbaijani language and, unlike other dictionaries, automatically determine the input language. The main drawback is that since example sentences were taken from the unedited parallel corpus, some errors occur. The project used to get financing from the Ministry of Communications. Currently the funding is scaled down.
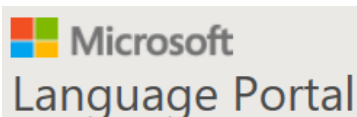
 Microsoft Language Portal belongs to Microsoft Corporation, which was one of the first to enter Azerbaijan software market. It is easy to guess that the dictionary includes words used for translating Microsoft products into Azerbaijani and mostly contains IT terminology.

 AzerDict dictionary is quite popular among existing electronic dictionaries. The website specifies 150,000 entries for English-Azeri language pair. Similarly to Polyglot, it is based on paper dictionaries dating back to 1994, 1998 and 2000. The orthographical dictionary includes 80,000 words while there is a new one for 100,000 words published long ago. The website's strength is its interactive feature – users can ask a question from the translation community or add words to the database. The dictionary is presumably updated based on statistics of most frequently searched results. However, during the 10 days spent writing this Policy Brief no new word was added.

 The AZLEKS project was developed in 2013 by "Altun kitab" publishing house and envisaged an

edition of both electronic and hard copy dictionaries. The project defined three priority areas:

- Terminology dictionaries
- Explanatory dictionaries
- Bilingual translation dictionaries.

According to the number of entries on the website, the focus has been made on explanatory dictionaries so far. Online dictionary has been prepared with support of K Dictionaries LTD an Israeli company that issued dictionaries in 50 world languages. Technically the website is well done; its weakness is that the linguistic section has only 30,000 words in English-Azeri and almost nothing vice versa.

Options

In order to improve the current offer of dictionaries there are 11 key areas that the publishing industry in Azerbaijan needs to focus on:

1. UPDATING. The dictionary should be regularly updated and reflect current changes in the language
2. MODERN sources. The lexicon should not be based on outdated sources
3. COMPLETENESS. It should have maximum number of entries including both words and phrases.
4. COMPREHENSIVE SEARCH RESULTS. It should show synonyms and antonyms, an example of use in sentences, grammar information and transcription. It should also include pictures illustrating the words where applicable.
5. TERMINOLOGY. It should contain special terminology on various subjects (medicine, finance, law etc.) and show search results by subject. For example, the meaning of the word vessel in medical and engineering fields is different.
6. HIGH QUALITY CONTENT. Content should be always created and edited by people and publishers should not rely entirely on translation software.
7. AN EFFECTIVE SEARCH ENGINE will ensure users can always find what they have searched for and more. Such tools as concordance (context) search, hyperlinks and word forms recognition are very useful.
8. USER FRIENDLY INTERFACE. A pleasant and convenient interface would include features that facilitate the use of an automatic switch for the input language, suggested corrections for mistyping and auto completion of words, among others.
9. INTERACTIVITY. The dictionary should not remain static. It should

respond to the needs of its audience, monitor search trends, provide a platform for information exchange among the translation community, allow the audience to add and comment on content.

10. ACCESSIBILITY. The dictionary should be well-promoted and widely-known. Desktop versions have to be available for download from a website and easily installed.

11. USING OTHER DICTIONARIES. It can include various dictionaries (other languages, explanatory, spelling, idiomatic, orthoepic dictionaries etc.).

Recommendations

The creation of a new dictionary would improve the strength of existing dictionaries and address their weaknesses and would be one step on the path that the government outlined for development of the Azerbaijani language. However, I suggest considering developing the dictionary within the broader context of establishing a national movement for ensuring the purity and beauty of the Azerbaijani language, called *Our Language Is a Treasure*. This approach would increase the dictionary's effectiveness.

The *Our Language Is a Treasure* movement would involve students, translators, professors of linguistics, copywriters, journalists, editors, writers, poets and other members of the linguistic community i.e. everybody that deals with a particular word, who could act as *language decipherers*. The broad promotion of the movement concept implies active work with students and various youth organizations. Young people have huge energy and an inherent need for belonging and self-actualisation. An attribute of membership to the movement would be a bracelet or emblem, which would allow members to recognize each other easily.

The older generation people also need to be involved in the movement – especially reputable language experts, like scholars from the Institute of Linguistics under the Academy of Sciences, renowned Azerbaijani writers and journalists. These people would act as a peer review panel on the website. Information on their participation in the project would be published on the website with their photos to ensure maximum transparency. They would also take part in regular social events which would host discussions on translation, learning and the correct use of Azerbaijani and foreign languages.

The establishment of the single vocabulary database is a collective effort. Best practice shows that such projects are successful with an open editing component. This includes the following projects and organizations:

[Project Gutenberg](#) is a volunteer effort to digitize and archive cultural works, to "encourage the creation and distribution of eBooks". It was founded in 1971 by Michael S. Hart and is the oldest digital library. As of October 3, 2015, Project Gutenberg reached 50,000 items in its collection. Within the project volunteers' involvement covers the whole process starting from scanning, double proof-reading and uploading the text on website. Every day over 2,000 pages undergo this process.

[LibriVox](#) is a group of worldwide volunteers who read and record public domain texts creating audiobooks available for free download from their website.

[Multitran](#) is the best online dictionary in Ru.net, containing over 10 million words and expressions, is also developed on volunteering principles.

And of course the most vivid example – the largest-ever created encyclopedia - [Wikipedia](#) that includes 40 million entries is also an open edited resource.

Taking this principle, the general idea that would underline the *Our Language Is A Treasure* movement would be: **A Small Contribution Can Make a Big Difference.**

Legislative background

The Government of Azerbaijan has always realized the vital role of language for nation building (national identity) and informational exchange between nations and generations and therefore, always paid special attention to this topic. A variety of laws have been adopted to ensure the development of the Azerbaijani language as the state national language including:

1. Decree of the President of Azerbaijan Republic no.506 dated June 18, 2001 *On Improved Use Of The State Language*

2. Decree of the President of Azerbaijan Republic no.552 dated August 9, 2001 *On Proclaiming Day of the Azerbaijani Alphabet and Azerbaijani Language*

3. The Law of Azerbaijan Republic *On the State Language in the Republic of Azerbaijan*

4. Resolutions of the President of Azerbaijan Republic no.55 dated January 12, 2004 *On Publication Of The Azerbaijani National Encyclopaedia* and no.56 dated January 12, 2004 O*n Mass Publications In the Azerbaijani Language in Latin Alphabet* including priority list of books subject to re-publication (January 12, 2004)

5. Resolution of the President of Azerbaijan Republic *On Financing The Publication of Works of Outstanding Authors of World Literature in the Azerbaijani Language* (September 18, 2008)

Full list and texts of legal acts related to the Azerbaijani language since 1918 are available here: http://files.preslib.az/projects/remz/pdf_ru/atr_dil.pdf.

More recently in 2013 within the Development Concept "Azerbaijan 2020: Look Into the Future" President of Azerbaijan Republic approved the «State Programme on Use of Azerbaijani language in globalization environment in accordance with requirements of the era and Development of linguistics in

the country».

The Programme has five priorities:

- Improved legal framework
- Enhanced teaching of the Azerbaijani language
- Enhanced use of the Azerbaijani language
- Promotion of the Azerbaijani language in mass media
- Use of information and communication technologies in research and promotion of Azerbaijani language

The Swiss-funded monitoring report on the outcomes of the programme produced by Entrepreneurship Development Foundation NGO within the Discussion Platform To Promote Development In Azerbaijan assessed progress on the following actions envisaged within the State Programme.

Actions	Status
2013 - Reorganize the Institute of Linguistics and renovate its facilities and infrastructure	A number of efforts have been made to expand the staff and reorganize structure of the Institute
2013-2015 - Prepare proposals on improving legislation;	From 2013 through 2016 there were no relevant legal changes
2013-2014 - Develop «National Transliteration Standards».	As of November, 2016 the work continues
2013-2014 - establish a hotline to involve wide public development of the Azerbaijani language, its preservation and purity as well as prevention of the violation of literary language norms, lexical and grammar rules in advertisements, in catering facilities etc.	The hotline has been operational since May 2016

2015 - Prepare proposals on establishing Azerbaijani language courses at Azerbaijani diplomatic representations for foreigners, Diaspora and other interested parties	The cultural sections of the Committee on Relations with Diaspora created a website in 2013 for distance learning on subjects like Azerbaijani language, history, culture, literature and geography of Azerbaijan. Since 2014 the website has been inactive.
Control stickers and names of imported in Azerbaijan goods to ensure they contain information necessary for consumers in the Azerbaijani language	This is being done regularly
2013-2014 - Develop software to transliterate Azerbaijani alphabet to other alphabets based on national transliteration standards	The software is under development
2013-2015 - Create a single database on the Azerbaijani language as the state language	In progress
2013-2015 - Develop a national terminology information system based on international standards	In progress
2013-2020 - Prepare new and improved orthographical, explanatory, phraseological, terminological, translation, encyclopedia and frequency electronic dictionaries of Azerbaijani language	All electronic dictionaries developed since 2013 have been produced by private sector
2013-2016 Create an anti-plagiarism system in Azerbaijani language	In April 2016 meetings were held with relevant ministries, agencies, universities and Academy of Sciences

Full report is available to download here:

http://edf.az/ts_general/azl/layihe/Azerbaijan-2020/hesabatlar/medeni_irs.pdf.

About dictionaries

The first dictionaries emerged along with the writing system. Actually they were not dictionaries but so-called "glosses" i.e. notes in the margins of ancient manuscripts with explanations of unknown words. Ancient Sumerians already had glosses in 2500 BC and the first full-scale dictionary in the world was created in 300 BC in China and included 2,094 articles and 13,113 hieroglyphs.

Dictionary Users

Since then dictionaries have been compiled in almost every language that has an established writing system. Today everybody uses dictionaries. They are probably the most read books. A survey conducted in Britain in the late 1980s showed that 90% of households had at least one dictionary at home. It is more than any other book including recipe books (70%) or even the Bible (80%). (Source: Lecture by Boris Iomdin Dictionaries – Myths and Reality https://www.youtube.com/watch?v=ppB6x-aP8Jc)

The lexicographical boom resulted in a series of dictionaries published by the same publishing houses, for example,

- Random House Webster, Barnhart, American Heritage (USA)
- Oxford, Chambers, Collins, Hamlyn, Longman, Macmillan (UK)
- Le Robert, Larousse, Tresor (France)
- Duden (Germany)
- Dictionaries of the Academy of Sciences (USSR, Russia)

Why do people need dictionaries?

A dictionary tracks language. The main purpose of the language is to reflect reality as it is. Language is a mirror of our life. The richer it is, the

better and more comprehensively we can describe our internal and outer world. But it can be reflected in different ways. Great texts are based on a rich vocabulary.

- Pushkin's language – 25,000 words
- Shakespeare's language – 24,000 words (it is believed that 1,700 of them he invented himself)
- Ordinary people's language – 5,000-10,000 words
- In order to understand a foreign language and speak fluently in a new environment it is enough to know 3,000 – 4,000 words
- There is a universal Natural Semantic Meta-language (NSM) developed by Anna Wierzbicka from the University of Canberra. It consists of 60 words, which are basic primitive notions that help explain all notions in the world.

A dictionary helps to expand an active vocabulary and create great texts. The rule of thumb for writers is always to use dictionaries and thesaurus to *find the right word for the job.* One of the most effective ways to achieve a rich, expressive writing style is to strive to find the perfect words to express yourself.

If words are not included in dictionaries, they tend to be forgotten, replaced with foreign equivalents and gradually lost. Linguistic archaization is quite natural. For example, Macmillan dictionaries use a smart system of stars for that purpose. Stars indicate how frequently the words are used in the language.

★★★ - first and most frequently used 2,500 words.

★★ - words within 2,500 – 5,000 range.

★ - covers over 5,000 – 7,500 most frequently used words. Words without a star are seldom used.

Reality around us and our life constantly changes. Change is the only thing that remains constant. Language is a living substance, it follows world trends and has to change all the time accordingly. We need the words like "*Brexit*" and "*smartphone*" to move on. If new words are not introduced in the language, it becomes obsolete and can eventually die. A dictionary is crucial in reflecting the metamorphosis of language.

A dictionary is a kind of legislator of the language. It is a reference point

to verify words. It sets standards and rules.

A dictionary tracks culture. Culture can be encompassed in the dictionary and preserved. Museums preserve world masterpieces of art, and dictionaries are responsible for preserving the language for future generations.

Main En-Az-En dictionaries (July 2017)					
	Polyglot	**Azerdict**	**Dilmanc**	**Microsoft**	**Azleks**
www	polyglot.az	azerdict.com	dilmanc.az	microsoft.com	azleks.az
1. Regular update	No	Yes	Stopped	No	Yes
2. Year of paper dictionary	En-Az 1994&2000 Az-En 1998	En-Az 1994&2000 Az-En 1998	n/a	n/a	No info
3. Number of entries	En-Az 23.500 Az-En 45.000	Az-En-Az 150.000	No info	4.200	En-Az 26.209
Search by phrases	No. Only add info to words.	Yes	No	Yes	Yes
4. Comprehensive search results					
Synonym/antonym	No	No	Ru-Az only	No	No
Sample of use in sentences	Seldom	No	Internet search results	No	Yes
Grammar	No	Yes	Limited	No	Yes
Transcription	No	No	No	No	Yes
5. Terminology	No	Limited	No	IT only	No
6. Quality content	Human origin	Human origin	Machine origin	Machine origin	Human origin
7. Effective search engine					
Concordance	Not workable	No	Internet search results	No	Yes
Hyperlinks	No	Yes	No	No	Yes
Word forms	No	No	Yes	No	No
8. User friendly interface					
Auto switching language	No	No	Yes	No	Yes
Offering options when wrong word inserted	No	No	No	No	Yes
Words auto completion	No	No	No	No	No
9. Interactivity	No	Yes	No	No	Only comments
10. Accessibility/ Popularity	Yes	Yes	Yes	No	Not very much
11. Other dictionaries	German, French, Russian	Russian, Orthography, Explanatory dictionary of Azeri language	Russian, Turkish	115 world languages	Explanatory, Spelling dictionary of Azeri language

Disclaimer

References

- http://www.unesco.org
- ru.wikipedia.org
- https://www.gutenberg.org/
- https://librivox.org
- https://www.multitran.ru/
- https://en.wikipedia.org
- http://preslib.az
- http://www.president.az
- http://edf.az
- https://academy.yandex.ru
- polyglot.az
- azerdict.com
- microsoft.com
- azleks.az
- dilmanc.az