Democratising AI: Polysemy in Azerbaijani NLP

Rauf Allahverdiyev
230945216
Dr. Haim Dubossarsky
MSc Computer Science
23 August 2024

***Abstract*—This research implies the challenging task of Natural Language Processing (NLP) in low-resource languages such as Azerbaijani despite having an average of 25 million speakers worldwide. The paper focuses on different language models differentiating a list of Azerbaijani polysemous words according to their meanings in context. Experimenting clusters with K-Means, the project focuses on the accuracy of three different models: XL-LEXEME, XLM-RoBERTa, BERT-TURKISH. Ultimately being chosen as the most promising model, XL_LEXEME performed an average of 0.61 F1 score and 0.62 accuracy score. This was carried out on 60 polysemous words from online Azerbaijani polysemous words dictionary and 4326 sentences from online Azerbaijani language corpus, "azcorpus". For the endeavours in future, this project will serve for transfer learning to be further refined, providing annotated phrases, a candidate list of polysemous words, integrated model, editable code, and other valuable insights regarding the NLP problem of other widely spoken but low-resource languages like Azerbaijani.**

Keywords*— *Azerbaijani, Polysemy, XL-LEXEME, AzCorpus, F1 score, Accuracy, XLM-RoBERTa, Model Fine-Tuning, Transfer Learning

I. Introduction

Oldest and eternal problems of the world have quickly shown its derivatives in 21st century life with the quick adaptation of technology into our daily routine. This paper will discuss one of these problems which is regarding injustice in technology, especially in artificial intelligence (AI). There have been calls to mitigate these problems by democratising AI, but these calls may not explore the aspects of a democratic AI to the fullest extent. According to Lin, there are three notable notions of "democratising AI": democratising AI use, democratising AI development, and democratising AI governance (Lin 2024). As Lin (2024) explains, the first notion is about how accessible AI is, the second notion is about how much a wider range of people is involved in the development of AI, and the last notion concerns regulations on use, access etc. This paper concentrates on the second concept more.

In most cases, low resource languages are not taken into account during AI development. Due to the lack of large corpora, pre-trained models, and linguistic databases, success in machine translation, sentiment analysis, speech recognition etc. can barely be achieved. Understanding this is not an easy task due to reasons such as unique scripts, dialectal variations, distinctive grammatical structures, it should be acknowledged that improvement in this area will provide a huge spectrum of benefit in an immense number of areas. Reference [2] is regarding how the speech market in Azerbaijan would benefit from AI use. Fatullayev, et al. (2021) emphasises the increasing number of calls received by the call center of the Ministry of Economy during the years 2016-2020 where the difference was threefold. Having an average of 3.9 minutes of talk over the span of five years, the total duration of talk time was 104500 hours. Obviously, this requires a huge number of call center staff and comes at a high cost. Implementation of speech technologies would reduce the cost of maintaining the organisation itself (Fatullayev, et al. 2021).

Authors discussed three reasons for the lack of the speech technologies development in Azerbaijani: lack of support for the Azerbaijani language, lack of security guarantees, low speech recognition accuracy. They do not find it extremely challenging to solve the first two problems since the development process is similar in Azerbaijani as other languages and security problems can be avoided by having local applications. However, the general handicap remains and the only way to overcome it is to create high-quality datasets and training models which would benefit from these datasets. In the end, the results will be of the similar quality to that of rich-resource languages.

Since polysemy is a fundamental part of natural languages and daily life, being able to successfully differentiate the meanings of polysemous words in a given context is a fundamental skill of human language comprehension. In this context, this research's main mission is to narrow down the selection of the appropriate model to further finetune. I achieve that by having tried several pre-trained models, calculated F1 and accuracy scores, and done comparisons in order to be able to suggest one of the models. With the help of my statistical data, some models can easily be eliminated at the beginning of the fine tuning process.

To sum up, the rest of the thesis will cover the origin of Azerbaijani language, unique features that impact the development process, the methods and resources I have used during research. In the second half, the thesis will focus on the developed code, its outputs and analysis of the result.

II. Resources

This thesis stems from several resources such as online polysemous words dictionary, language corpus, existing ML models etc. Initially, I referred to [3] for the dictionary of all polysemous words in Azerbaijani. After scraping the website, I could use the words, their part of speech, and definitions for each sense.

Later, to look through a corpus, I requested access to use [4]. This is an Azerbaijani language corpus created by Kishiyev H. et al. Named "azcorpus", this corpus contains 1.9 million documents to be used to generate text. In more depth, "azcorpus" benefits from mainly three sources:

1. az_books: 1,540,732 instances (19 GB)
2. az_wiki: 98,882 instances (0.9 GB)
3. az_news: 236_878 instances (3.8 GB)

I took advantage of “azcorpus” to manually annotate sentences which would be fed to the three models I used to evaluate the best one. The models are:

1. **XLM-RoBERTa** is a pretrained model on 100 languages. The model was trained on 2.5 TB of filtered CommonCrawl data, while mBERT was trained on Wikipedia data. It outperformed mBERT in cross-lingual benchmarks including XNLI, MLQA, and NER. (Dubossarsky, H. and Dairkee, F. 2024)
2. **BERT-TURKISH** is an uncased BERT model for Turkish created and supported by the community. The current model is trained on a filtered and sentence segmented version of the Turkish OSCAR corpus, a recent Wikipedia dump, many OPUS corpora, and a special corpus donated by Kemal Oflazer. The final training corpus is 35GB in size and includes 44,04,976,662 tokens [11].
3. **XL-LEXEME** is a model that applies pairwise sequence similarity models to the WiC challenge, highlighting the target word for LSC detection [5].

Finally, this thesis used a variety of materials and approaches to address the difficulty of understanding polysemy in the Azerbaijani language. The first stage was to use a specialised online dictionary [3] to collect polysemous terms and their related meanings, which served as the foundation for the linguistic dataset. The following use of the "azcorpus" [4] substantially enhanced the research, providing a comprehensive collection of Azerbaijani texts that enabled manual annotation for model evaluation. Using advanced models like XLM-RoBERTa, BERT-TURKISH, and XL-LEXEME, this study was able to systematically examine the effectiveness of various techniques in the context of Azerbaijani polysemy. The successful integration of various resources emphasises the significance of a multifaceted strategy to addressing linguistic difficulties in low-resource languages.

.

III. Azerbaijani Language

Azerbaijani (Azeri in some other contexts), is a Turkic language spoken predominantly in Azerbaijan as well as in the neighbourhoods such as Iran, Georgia, Russia, and Turkey. The language has about 23-30 million speakers globally, with the majority living in Azerbaijan and northwestern Iran. Azerbaijani belongs to the Oghuz branch of the Turkic language family, which additionally includes Turkish, Turkmen, and Gagauz. Historically, the language has been influenced by Persian, Arabic, and Russian during various eras of governmental and cultural dominance in the region. The language has two main dialects: Northern Azerbaijani (spoken in Azerbaijan) and Southern Azerbaijani (spoken in Iran). After the Soviet Union's disintegration, Northern Azerbaijani shifted from using the Cyrillic script to a Latin-based alphabet, whereas Southern Azerbaijani still uses a variant of the Persian script.

When it comes to the unique features of Azerbaijani, it has a rich agglutinative morphology using affixes to describe grammatical relationships and semantic nuances. This is a frequently observed element of Turkic languages, allowing for the formation of complicated words with several suffixes expressing tense, mood, person, and case. For instance, a single Azerbaijani verb can have signs for subject agreement, negation, and tense all in the same word.

Moreover, Mokari, P.G. and Werner, S. (2017) discusses prosodic features of Azerbaijani in his article. The syllables in Azerbaijani are V, VC, CV, or CVC. In Azerbaijani, stress is typically placed on the final syllable of a word, with the exception of imperative verbs and negative suffixes where the first syllable is emphasised. Southern dialects use rising intonation instead of interrogative particles to mark polar queries, influenced by Persian. Declaratives and wh-questions typically have a lower pitch at the end. Sentences with multiple phrases typically end with a rising pitch, save for the final phrase. The pitch gradually decreases across successive phrases. In yes-no questions, there is no final fall and the last syllable is pitched higher than others [12].

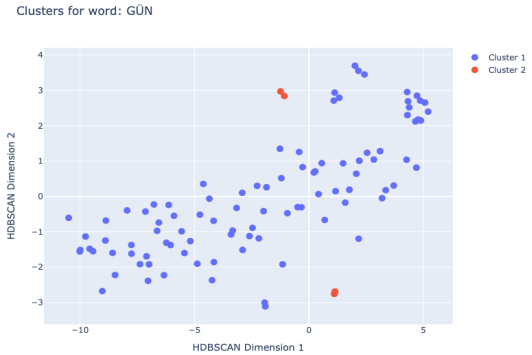
Another distinguishing element of Azerbaijani is vowel harmony, a phonological process in which vowels within a word harmonise to become front or back vowels, depending on the root vowel. This harmony promotes phonological coherence within the word and is an essential component of Azerbaijani phonology. Furthermore, Azerbaijani also includes a case system, however it is less comprehensive than in other inflectional languages. There are six cases: nominative, accusative, dative, genitive, locative, and ablative, each with a different syntactic role in a sentence. The language's evolution through several writing systems, notably in the last century, demonstrates its adaptability and the impact of political changes on linguistic patterns. Lastly, Azerbaijani does not have a gender definition for the third person pronouns which is the case for Russian, English and so on.

IV. Methods

In terms of the chronology of the project, I started with sorting all polysemous words I scraped from [3] using the Python package “Beautiful Soup” [8] by their frequency in “azcorpus”. Afterwards, some manual exclusion was done, only keeping 60 of them which showed polysemy across only nouns.

The next step was to collect 80 sentences for each word (40 sentences per sense) and to choose the best model to move forward. In that manner, I initially fetched 20 sentences for the first 6 words from “azcorpus” and manually annotated them, creating their true labels as 1 (dominant sense) or 0 (subordinate sense). After plotting them with all three models, I was confident about excluding BERT-TURKISH due to its poor performance. Although I initially hoped for a good result because of the similarities between Azerbaijani and Turkish, this was not the case for the models.

Fig. 1. “Gün (Day(1) / Sun (0))” plotted with BERT-TURKISH



After considering only XL-LEXEME and XLM-RoBERTa [9], their plots showed roughly the same performance. Thus, I had to calculate F1 and Accuracy scores in order to make a more concise decision. All F1 scores were calculated using the Scikit package [8].

	XL-LEXEME		XLM-RoBERTa	
	F1	Accuracy	F1	Accuracy
Dövlət	0.88	0.79	0.82	0.7
Gün	0.6	0.6	0.56	0.55
Qeyd	0.74	0.59	0.91	0.85
Davam	0.71	0.55	0.82	0.7
Qaz	0.7	0.63	0.68	0.5
Məsələ	0.55	0.61	0.68	0.6
Averages	0.7	0.62	0.73	0.66

Table 1: F1 and Accuracy scores of the first 6 words plotted with models XL-LEXEME and XLM-RoBERTa

Since this is data for only a small sample and there is not much difference between averages, I decided to move forward with XL-LEXEME. Afterwards, I did the same procedure for 1000 sentences per word. Among approximately 60,000 sentences, I annotated some of them manually, and some of them using the plots aiming to end up with approximately 80 sentences for each of the 60 words. Regarding the work principle of XL-LEXEME, it calculates the Levenshtein distance between two strings and finds the start and end positions of the word in the sentence with the smallest Levenshtein distance. This is the final statistics of the words and sentences:

Total Number of Words	60
Total Number of Sentences	4,326
Average Number of Sentences per Word	72
Average Number of Sentences per Sense	36

Table 2: Summary statistics of words and sentences collected

V. Results & Analysis

This section starts with what the online dictionary looks like. Reference [3] contains words, their part of speech, origin, sense meaning, and one example for each sense. The page would look like this:

1. *Dövlət, is. [ər.] - Hökümət*
2. *Dövlət, is. [ər.] - Əmlak, sərvət. 'Dövlət ki, yerin deyişindən çıxmaz.'*

The plots for the first 5 words are like this:

1.

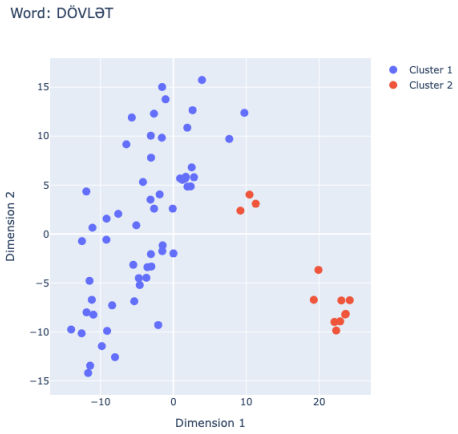


Fig. 2: Dövlət - Government (dominant/1) / wealth (subordinate/0)

As it is obvious from the plots, XL-Lexeme did a great job differentiating these two senses into clusters using K-means. The F1 and accuracy scores are also decent, being 0.64 and 0.68, respectively.

2.

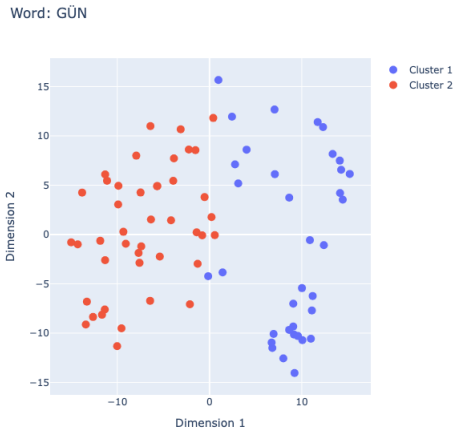


Fig. 3: Gün - Day (1) / Sun (0)

XL-Lexeme has shown a bit poorer performance here, since F1 and accuracy scores are lower, being 0.51 for both.

3.

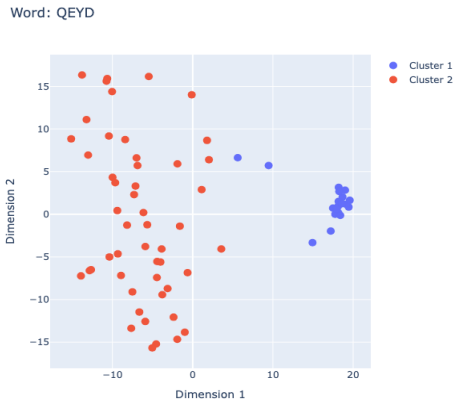


Fig. 4: Qeyd - Note (1) / Care (0)

The model has done an excellent job for “Qeyd”. F1: 0.74; Accuracy: 0.75. This is much higher than the average of 60 words, which is F1: 0.61; Accuracy: 0.62

4.

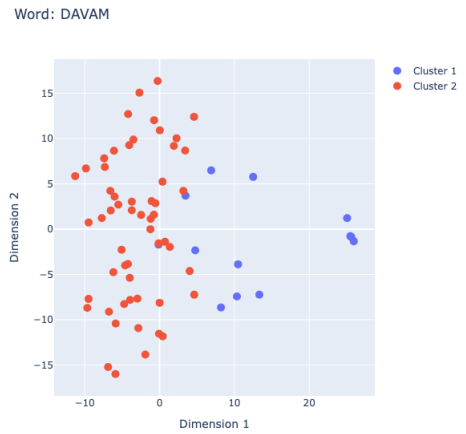


Fig. 5: Davam- Continuation (1) / Resistance (0)

The scores are a bit short of the average, but decent job in general. F1: 0.53; Accuracy: 0.58

5.

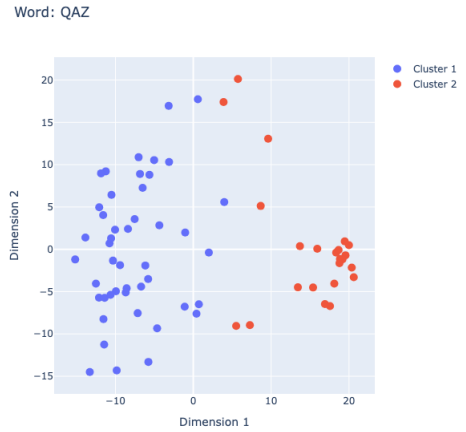


Fig. 6: Qaz - Gas (1) / Geese (0)

Another excellent job by the model. The scattered dots signify high F1 and accuracy scores, which are 0.67 and 0.68, respectively.

These 5 examples proved the ability of the model to differentiate the meanings and put them into clusters. The hovering functionality assisted me to annotate some of the sentences for the rest of the candidate list. Hovering example is shown below:

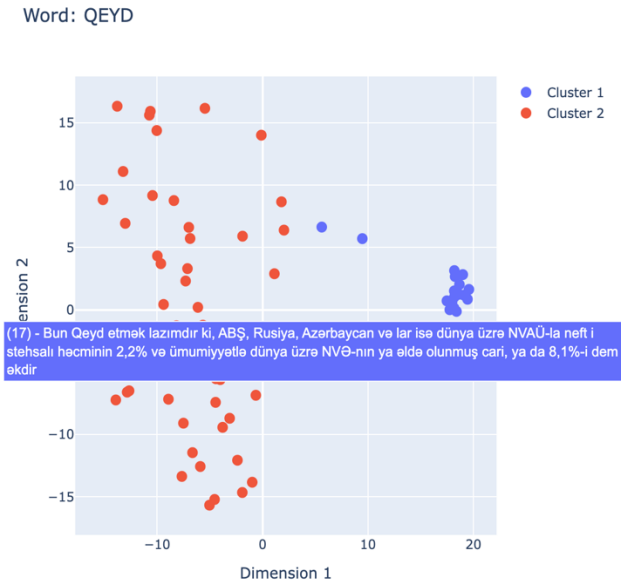


Fig. 7: Hovering function for “Qeyd”

VI. Discussion & Conclusion

This part of the thesis will discuss the results of the project, referring to the plots, F1 and accuracy scores etc. Getting an average of 0.61 F1 score and 0.62 accuracy score can be considered a satisfactory result. However, this still leaves room for improvement. In that manner, there are several ways for a better F1 and accuracy score.

The path starts with data augmentation. Expanding the dataset as much as possible will enable the model to understand senses in a variety of contexts and hence will improve generalisation. This should be followed by feature engineering. Adding part of speech information and surrounding the target sentence with context supplementary sentences will enhance the model. The next step is to fine-tune a model with more domain-specific data. Experimenting with different learning rates, batch sizes, and other hyperparameters will maximise the model performance. As it has been done at the earlier stages of this project, a mix of several models will also play a crucial role in model selection. Conducting a thorough error analysis to uncover prevalent misclassifications and relabeling unclear or wrongly marked instances will add more sophistication to the process. Furthermore, cross-validation ensures that the model's performance is consistent across different data subsets. In other words, this reduces overfitting and yields a more trustworthy assessment of accuracy and F1 score. Lastly, adjusting the decision threshold for classification for a better balance precision and recall is vital, particularly in circumstances where the dataset is uneven. By following these steps consistently, XL-LEXEME's performance on the annotated phrases is very likely to improve, resulting in an improved accuracy and F1 scores.

In conclusion, there is still a way to go further in this area. A very serious commitment should be made to bring Azerbaijani closer to the rich-resource language level. This work provides a pack of things as a start point and will help

transfer learning with its model, candidate list of words, annotated sentences and other resources used during the preparations of this work.

REFERENCES

[1] Lin, Ta. (2024) "Democratizing AI' and the Concern of Algorithmic Injustice.", Philosophy & Technology, 37(3). Available at: https://doi.org/10.1007/s13347-024-00792-2

[2] Abbasov, A., Fatullayev, A. and Fatullayev, R. (2021) "Speech Technologies Market in Azerbaijan", American Journal of Management,21(3).Availableat: https://doi.org/10.33423/ajm.v21i3.4365

[3] Azərbaycan Dilinin Omonimlər Lüğəti. Available at: https://obastan.com/azerbaycan-dilinin-omonimler-lugeti

[4] Kishiyev, H. et al. Azcorpus/azcorpus_v0 · datasets at hugging face, azcorpus - The largest open-source NLP corpus for Azerbaijani . Available at: https://huggingface.co/datasets/azcorpus/azcorpus_v0

[5] Cassotti, P. et al. (2023) 'XL-Lexeme: WIC pretrained model for cross-lingual lexical semantic change', Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) [Preprint]. Available at: https://doi.org/10.18653/v1/2023.acl-short.135

[6] Cassotti, P. et al. (2023) Pierluigic/XL-Lexeme: WIC pretrained model for cross-lingual lexical semantic change, GitHub. Available at: https://github.com/pierluigic/xl-lexeme?tab=readme-ov-file

[7] F1_score, scikit. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

[8] Beautiful Soup Documentation¶ Beautiful Soup Documentation - Beautiful Soup 4.12.0 documentation. Available at: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[9] XLM-Roberta (no date) XLM-RoBERTa. Available at: https://huggingface.co/docs/transformers/en/model_doc/xlm-roberta

[10] Dubossarsky, H. and Dairkee F. (2024). Strengthening the WiC: New Polysemy Dataset in Hindi and Lack of Cross Lingual Transfer. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING2024).Availableat: https://aclanthology.org/2024.lrec-main.1332.pdf

[11] Dbmdz/bert-base-turkish-128k-uncased · hugging face dbmdz/bert-base-turkish-128k-uncased · Hugging Face. Available at: https://huggingface.co/dbmdz/bert-base-turkish-128k-uncased

[12] Mokari, P.G. and Werner, S. (2017) 'Azerbaijani', Journal of the International Phonetic Association, 47(2), pp. 207–212. Available at: https://doi.org/10.1017/S0025100317000184

APPENDIX

[a]

word	translation	frequency	PoS-1	PoS-2
DÖVLƏT	State (1) / Wealth (0)	62973	Noun	Noun
GÜN	Day (1) / Sun (0)	41874	Noun	Noun
QEYD	Note (1) / Care (0)	40200	Noun	Noun
DAVAM	Continuation (1) / Resistance (0)	23983	Noun	Noun
QAZ	Gas (1) / Geese (0)	8760	Noun	Noun
MƏSƏLƏ	Issue (1) / Math Exercise (0)	8027	Noun	Noun
MADDƏ	Constitution principle (1) / Substance (0)	5710	Noun	Noun
DAĞ	Mountain (1) / Grief (0)	5423	Noun	Noun
ÇAY	River (1) / Tea (0)	4709	Noun	Noun
BAZAR	Market (1) / Sunday (0)	4585	Noun	Noun
DÖVR	Time period (1) / Cycle (0)	3928	Noun	Noun
ƏSƏR	Art/Literature Work (1) / Trace (0)	3729	Noun	Noun
GÖRÜŞ	Meeting (1) / Opinion (0)	3444	Noun	Noun
MAL	Cattle (1) / Commodity (0)	2850	Noun	Noun
TON	Ton (1) / Tone (0)	2784	Noun	Noun
FORMA	Shape (1) / Uniform (0)	2324	Noun	Noun
QOL	Goal (1) / Arm (0)	2128	Noun	Noun
ATƏŞ	Gunfire (1) / Fire Pit (0)	2052	Noun	Noun
TOP	Gun (1) / Ball (0)	1904	Noun	Noun
TAXTA	Wood (1) / Blackboard (0)	1875	Noun	Noun
BAL	Honey (1) / Point (0)	1731	Noun	Noun
BORC	Debt (1) / Duty (0)	1717	Noun	Noun
BAĞ	Garden (1) / Tie (0)	1642	Noun	Noun
QOYUN	Sheep (1) / Lap (0)	1523	Noun	Noun
PAY	Gift (1) / Share (0)	1412	Noun	Noun
ŞƏKƏR	Sugar (1) / Diabetes (0)	1351	Noun	Noun
ŞAM	Pine (1) / Candle (0)	1351	Noun	Noun
PARÇA	Bit (1) / Fabric (0)	1280	Noun	Noun
XAL	Score (1) / Freckles (0)	1195	Noun	Noun
GÜNLÜK	Daily (1) / Hat (0)	1101	Noun	Noun
MAYA	Capital (1) / Yeast (0)	1078	Noun	Noun
KÜTLƏ	Large piece (1) / Mass (0)	1038	Noun	Noun
ŞƏR	Evil (1) / Night (0)	942	Noun	Noun
QURD	Wolf (1) / Worm (0)	923	Noun	Noun
BEL	Waist (1) / Shovel (0)	819	Noun	Noun
BÖLMƏ	Section (1) / Division (0)	726	Noun	Noun
OCAQ	Hearth (1) / Home (0)	702	Noun	Noun
OBYEKT	Property (1) / Object (0)	673	Noun	Noun
AKT	Deed (1) / Event place (0)	637	Noun	Noun
DAVA	Fight (1) / Medicine, Cure (0)	637	Noun	Noun
BULUD	Cloud (1) / Wide plate (0)	583	Noun	Noun
MAT	Shocked (1) / Checkmate (0)	531	Noun	Noun

XƏRÇƏNG	Disease (1) / Cancer (0)	514	Noun	Noun
PARA	Piece (1) / Money (0)	466	Noun	Noun
QAYNAQ	Welding (1) / Source (0)	451	Noun	Noun
ŞTAT	State (1) / Staff (0)	395	Noun	Noun
BAS	Press (1) / Bass (0)	363	Noun	Noun
TUŞ	Faced (1) / Mascara (0)	360	Noun	Noun
ŞAN	Glory (1) / Beehive (0)	341	Noun	Noun
KÜRƏ	Sphere (1) / Earth (0)	338	Noun	Noun
BOĞAZ	Throat (1) / Bosphorus (0)	328	Noun	Noun
APARAT	Device (1) / Government staff (0)	323	Noun	Noun
ÇANAQ	Pelvis (1) / Musical instrument part (0)	270	Noun	Noun
AVAR	Paddle (1) / Folk name (0)	266	Noun	Noun
BAN	Chassis (1) / Rooster crow (0)	251	Noun	Noun
VURĞU	Emphasis (1) / Beat (0)	249	Noun	Noun
DÜYÜN	Knot (1) / Wedding (0)	237	Noun	Noun
QAŞ	Eyebrow (1) / Precious stone (0)	236	Noun	Noun
ƏQRƏB	Scorpion (1) / Hand of clock (0)	198	Noun	Noun
KÜRƏK	Back (1) / Paddle (0)	129	Noun	Noun