NATURAL LANGUAGE PROCESSING (CSCI-6515 - 20543)

Project #1

Authors: Nijat Jafarov, Namig Planov

Folktales in Azerbaijani Dataset

Source: Azerbaijani WikiSource

License: Public Domain

id	title	text	source_url
1	Ac qurdun na…	Biri var idi, biri yox idi, bir qurd var idi. Bu qurd mahalda qoyun, keçi qoymurdu, yeyirdi. Amma yenə də ac idi. Günlərin bir günü şirlə pələng …	https://az.…
2	Adamcıl	Bir çoban axşam vaxdı sürünü gətirirmiş kəndə. O, qəbiristanlıqdan keçəndə görür ki, bir iri buynuzdarı və uzun caynaqları əcayib bir qəbr eş…	https://az.…
3	Altı dul arvad	Bir gün altı dul arvad bir yerdə toplaşdılar. Bunların biri təklif edir ki, gəlin hər birimiz öz başına gələni söyləsin. Arvadlar bu təklifə razı oldular…	https://az.…
4	Altı yoldaş	Badi-badi giriftar, hamam-hamam içində, xəlbir saman içində, dəvə dəlləklik elər, köhnə hamam içində. Qarışqa şıllaq atdı, dəvənin qıçı batdı…	https://az.…
5	Armud bəy	Biri vardı, biri yoxdu, bir tülkü vardı. Bu tülkü bir gün girrənə-girrənə gedirdikdə görür ki, bir armud ağacı var. Ağacın altı doludur armudlarla: y…	https://az.…
6	Avçı Məhəm…	Günlərin bir günündə əyyami sabiqdə bir danə Məhəmməd var idi. Günlərin bir günündə Məhəmməd çöldə yer şumlayırdı. Məhəmməd qan-t…	https://az.…
7	Axvay	Gülnahar mahalında Şəftən adlı orta sinli bir kişinin Gülxar adlı bir arvadı var idi. Şəftənin peşəsi naxırçılıq, Gülxarın peşəsi evdarlıq idi. Bu ə…	https://az.…
8	Ağ quşun nağılı	Biri vardı, biri yoxdu, bir padşah vardı. Bütün padşahlardan fərqli olaraq, bu padşah olduqca adil və ağıllı padşah idi. Hamı onu çox sevirdi, bi…	https://az.…
9	Ağıllı qız	Həkan-həkan içində, qoz girdəkan içində, dəvə dəlləklik eylər, köhnə hamam içində. Biri varıydı, biri yoxuydu, bir padşahnan vəzir varıydı. P…	https://az.…
10	Ağıllı qoca	Biri var imiş, biri yox imiş, uzaq keçmişlərdə qəddar bir padşah var imiş. Bu şahın qoyduğu qanuna görə övladlar əldən düşmüş qoca ata-an…	https://az.…
11	Bacı-qardaş il…	Biri varmış, biri yoxmuş, bir qoca kişi varmış, onun bir qızı, bir oğlu varmış. Qoca hər gün gedir üç qardaş divin anbarından bir dağarcıq darı …	https://az.…
12	Baftaçı Şah A…	Günlərin bir günündə Şah Abbas vəzirini çağırıb dedi: – Vəzir, neçə aydı ki, seyrə çıxmamışam. Qoşun hazırlığı gör, tədarük elə, on günün s…	https://az.…
13	Barxudarın b…	Biri var idi, biri yox idi. Yaradan bir idi, bəndəsi çox idi. Barxudar adlı bir balıqçı var idi. Barxudar dəryadan balıq dutar, bazarda satıb dolanar …	https://az.…
14	Barxudarın n…	Biri var idi, biri yox idi, yaradan var idi, bəndəsi yox idi. Barxudar adlı bir balıqçı vardı. Barxudar dəryadan balıq tutar, bazarda satıb dolanardı…	https://az.…
15	Beçə Dərviş	Raviyani əxbar, nağılan asar, şirin şəkkər, xoş göftar. Dedim getmə qal, Gəl ol abdal, Mən olum diləfkar. Dilimdən nə dedim yarə, Eşq odunu …	https://az.…

Datasheet for the Dataset

(Gebru et al., 2021)

Motivation: Purpose, creators, funding

Context: How and when texts were produced

Language: Language and dialect

Authors: Known demographics

Data: Size, sampling, preprocessing

Annotations: Type and process

License: Copyright and usage

Datasheet for Folktales in Azerbaijani Dataset

Nijat Jafarov, Namig Planov

January 2026

1 Datasheets for datasets

In this work, we adopt the Datasheets for Datasets framework proposed by Gebru et al. (2020) to document an Azerbaijani-language text corpus composed of folktales The dataset is collected from Azerbaijani Wikisource, which hosts texts that are either in the public domain or available under the Creative Commons Attribution-ShareAlike (CC BY-SA) license.

Given the scarcity of well-documented Azerbaijani NLP datasets, especially for low-resource language research, this datasheet aims to clearly describe the motivation, composition, collection process, licensing, and intended uses of the dataset. This documentation is intended to support transparency, reproducibility, and responsible reuse of the dataset in academic research.

2 Template

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

This dataset was created for academic research and university-level projects in the area of Natural Language Processing (NLP). The primary motivation was to support experimentation, analysis, and model development for Azerbaijani.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

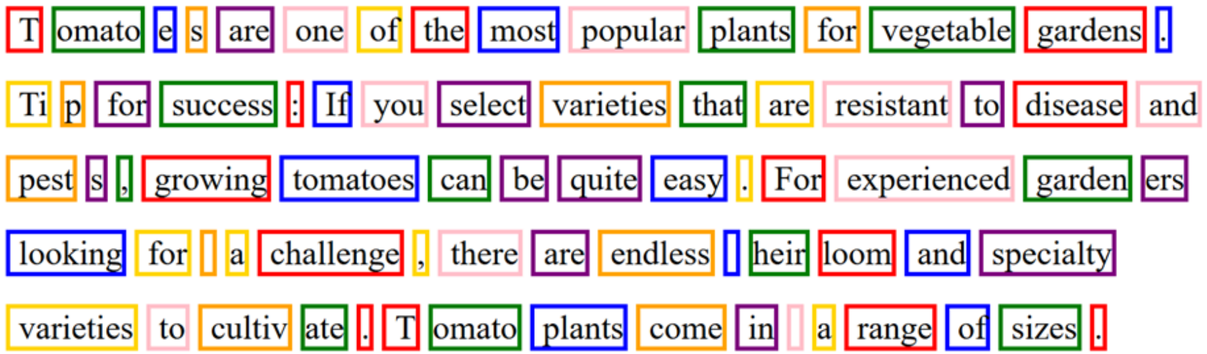
This dataset was created by two graduate students, Nijat Jafarov and Namig Planov enrolled in an NLP course at ADA University and The George Washington University. The dataset was developed as part of a university course project and was not created on behalf of any commercial organization, company, or external institution.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

1

Tokenization

Tokenization involves dividing a Textual input into smaller units known as tokens. These tokens can be in the form of words, characters, sub-words, or sentences. It helps in improving interpretability of text by different models.



Number of tokens: 397689

Number of types: 40158

Different approaches to tokenization

Word: ["kitablar"]

Character: ["k", "i", "t", "a", "b", "l", "a", "r"]

Subword: ["kitab", "lar"]

N-Gram: ["kit", "ita", "tab", "abl", "bla", "lar"]

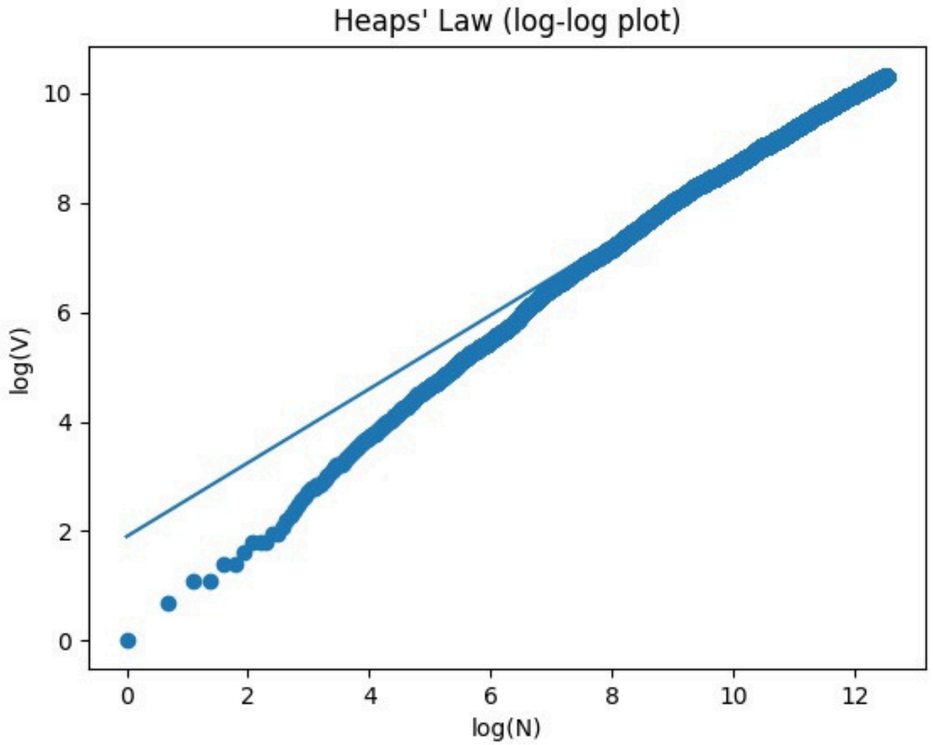
BERT Tokenizer: subwords, fixed vocab, OOV handling

BPE: frequent merges, data-driven, subwords

Unigram LM: probabilistic, multiple segmentations

SentencePiece: language-independent, raw text, subwords

Heaps’ law



Heap's law states that the number of unique words **V** in a collection with **N** words is approximately **Sqrt[N]**. The more general form of this law is:

In our case:

K: 7.175147430347003

Beta: 0.6666136782602733

For morphologically rich languages:

Beta : 0.5 - 0.75

K: 5 - 20

Byte Pair Encoding

It breaks down words into smaller, meaningful pieces called subwords. It works by repeatedly finding the most common pairs of characters in the text and combining them into a new subword until the vocabulary reaches a desired size. This technique helps in handling rare or unknown words by breaking them into smaller parts that the model has already learned during training. By reducing the vocabulary size, it makes it easier to work with large amounts of text while allowing the model to understand wide variety of languages.

Example:

('n', '</w>'), ('i', '</w>'), ('r', '</w>'), ('ə', '</w>'), ('a', '</w>'), ('ı', '</w>'), ('l', 'a'), ('d', 'i</w>'), ('b', '</w>'), ('b', 'i'), ('m', 'ə'), ('m', '</w>'), ('l', 'ə'), ('a', 'r'), ('u', '</w>'), ('ı', 'n'), ('d', 'ə'), ('d', 'a'), ('a', 'n'), ('bi', 'r</w>'), ('b', 'a'), ('i', 'n'), ('d', 'ı</w>'), ('g', 'ö'), ('ə', 'l'), ('d', 'ü'), ('i', 'r'), ...

Sentence Segmentation

Sentence segmentation is the process of determining the longer processing units consisting of one or more words. This task involves identifying sentence boundaries between words in different sentences.

Number of sentences in our case:

47, 578

Difficulties in Azerbaijani:

Prof. Əliyev dərsə gəldi.

"Gəlirəm", dedi.

Bəlkə də... bilmirəm.

— Haraya gedirsən? — Evə.

Mən gəldim sən getdin

Spelling Checker by Levenstein distance

Operations in Levenshtein distance are:

Insertion: Adding a character to string

Deletion: Removing a character from string.

Replacement: Replacing a character in string with another character.

Example:

gor**u** → gor (deletion of “u”)

kita**p** → kita**b** (substitution of “b” for p”)

səlam → səlam**ət** (insertion of “ə” and “t”)

levenshtein distances will be 1, 2, 2 accordingly

Confusion Matrix

correct \ typed	a	b	c	ç	d	e	ə	f	g	ğ	h	x	ı	i	j	k	q	l	m	n	o	ö	p	r	s	ş	t	u	ü	v	y	z
a	0	37	8	7	30	171	70	5	6	7	8	3	470	256	2	23	7	21	28	44	104	5	10	28	45	6	16	157	14	9	28	7
b	13	0	18	14	57	10	5	21	17	1	29	9	2	4	3	27	25	42	33	20	6	0	27	31	58	16	34	4	1	24	37	10
c	7	21	0	12	130	6	6	6	12	29	7	10	10	27	15	12	11	57	61	63	1	1	2	74	57	20	14	1	6	11	64	17
ç	9	12	35	0	161	3	5	4	5	10	2	7	19	30	2	13	7	41	41	60	2	1	4	11	71	12	23	8	1	15	40	6
d	9	33	31	18	0	13	9	19	30	18	15	16	21	38	3	39	18	57	158	72	11	3	21	33	151	17	72	5	7	19	75	12
e	42	1	3	0	10	0	18	0	1	0	3	1	9	44	0	1	2	6	3	1	11	10	5	3	1	3	2	14	3	0	4	1
ə	106	18	10	5	26	793	0	2	5	3	8	4	26	448	0	14	8	14	21	28	27	20	4	18	26	3	7	28	73	5	14	7
f	3	7	7	2	14	4	0	0	1	1	5	3	0	2	2	7	9	11	15	19	0	0	9	12	17	3	9	3	0	5	13	3
g	2	12	8	3	33	9	7	7	0	1	8	1	0	7	3	13	5	5	13	7	1	2	0	1	33	9	6	2	0	4	7	4
ğ	0	1	5	10	12	0	0	0	33	0	2	12	0	0	1	3	4	15	11	23	1	3	4	13	18	6	8	2	0	3	15	4
h	5	6	8	1	18	2	2	6	11	2	0	10	0	0	1	12	6	11	7	12	2	1	4	15	11	5	4	0	0	4	27	5
x	8	20	14	6	29	0	1	7	6	6	5	0	2	3	0	22	12	27	15	31	1	1	7	13	23	14	23	3	0	8	21	14
ı	210	1	0	5	10	12	7	1	2	1	3	2	0	256	1	5	4	11	18	10	14	1	0	4	4	1	5	11	3	2	4	1
i	93	5	3	4	27	159	266	4	5	1	12	2	34	0	3	12	4	13	20	15	36	7	3	14	19	1	11	32	18	0	27	2
j	1	2	0	1	0	0	0	0	2	0	0	1	0	1	0	0	2	2	2	4	0	0	0	2	0	1	2	0	0	3	5	1
k	15	52	53	13	80	19	13	18	28	5	31	12	2	14	9	0	18	43	89	126	12	5	14	145	40	21	41	9	1	12	187	41
q	23	73	20	17	50	12	3	16	100	227	25	32	4	13	2	78	0	35	143	144	10	3	18	175	82	9	22	2	3	21	55	28
l	21	186	46	26	820	25	22	15	20	55	13	17	74	128	5	96	38	0	265	438	9	2	17	289	307	38	53	23	25	40	204	58
m	23	34	33	13	204	12	11	10	21	27	16	11	37	40	2	38	24	42	0	65	6	1	10	118	294	18	35	9	3	12	118	17
n	18	96	36	11	104	18	9	23	16	29	17	28	6	15	1	89	62	79	331	0	5	1	20	457	78	34	45	7	5	26	112	43
o	82	6	1	1	5	39	13	2	4	1	4	0	16	42	1	6	4	6	6	6	0	3	7	5	10	2	5	28	3	0	4	1
ö	4	1	1	1	4	21	7	1	2	0	0	0	0	22	0	0	1	3	2	1	15	0	2	0	1	2	2	6	17	0	2	1
p	6	10	10	4	7	5	0	9	10	5	8	3	1	4	2	28	4	15	10	11	10	1	0	12	19	9	14	6	0	7	14	7
r	17	67	41	17	63	31	7	19	20	64	23	27	11	18	4	80	45	81	313	222	11	0	20	0	75	30	58	6	1	18	113	50
s	10	18	31	21	115	19	5	14	21	13	23	19	22	44	9	26	28	36	80	117	13	2	16	45	0	15	45	13	4	28	59	20
ş	13	131	14	11	35	9	1	15	8	21	8	17	1	6	1	28	21	45	117	193	2	0	7	357	72	0	30	1	1	15	55	44
t	19	24	30	31	128	35	21	20	37	7	25	17	17	47	1	41	19	54	93	73	8	1	32	110	71	26	0	5	6	25	60	32
u	69	2	4	1	3	22	12	1	1	0	2	1	11	23	2	1	1	0	4	5	34	1	2	5	6	1	1	0	13	3	2	0
ü	4	0	1	1	7	37	44	1	1	0	0	0	1	19	0	1	1	5	2	0	9	9	3	0	4	1	3	52	0	0	0	0
v	6	8	4	2	17	1	3	9	8	1	6	2	2	3	0	9	9	44	12	21	4	0	7	18	34	3	8	3	1	0	19	4
y	7	36	28	20	115	7	2	11	29	22	14	15	2	11	2	27	31	39	56	60	9	1	16	68	139	23	31	6	0	24	0	32
z	5	10	17	8	45	6	4	6	8	21	7	9	2	5	1	28	23	43	62	220	1	2	11	63	42	20	15	2	2	3	26	0

Confusion Matrix, our approach

We found Azerbaijani dictionary file from **Mozilla Azerbaijani Spell Checker** and messy Azerbaijani Twitter data. Using bigrams, we quickly found which dictionary words were close to the messy tokens. Then, we measured exactly which letters were swapped, building a full confusion map of the alphabet.

Difficulties with misspelling:

yaxasından yaxından s m 72

yaxci yaxma m c 30

yaxcı yaxı m c 30

yaxdi yaxdır ı i 225

yaxdısa yaxdırt r s 64

yaxin yasin s x 17

yaxinlarda yaxınlarda ı i 226

THANK YOU