

بسم الله الرحمن الرحيم

پردازش زبانهای طبیعی

جلسه ۱

احسان الدین عسگری

بهمن ۱۴۰۲

<http://language.ml/>

دانشکده مهندسی کامپیوتر

آزمایشگاه پردازش هوشمند متن و زبان و علوم انسانی محاسباتی

asgari@berkeley.edu



— زبان



تعريف زبان

• معرفی

• زبان

• پردازش زبان طبیعی

• درس NLP

• جلسه بعد

Chomsky (1959: 137) “A language is a collection of sentences of finite length all constructed from a finite alphabet (or, where our concern is limited to syntax, a finite vocabulary) of symbols.”



تعريف زبان

• معرفی

• زبان

• پردازش زبان طبیعی

• درس NLP

• جلسه بعد

Chomsky (1959: 137) “A language is a collection of sentences of finite length all constructed from a finite alphabet (or, where our concern is limited to syntax, a finite vocabulary) of symbols.”



DNA Language

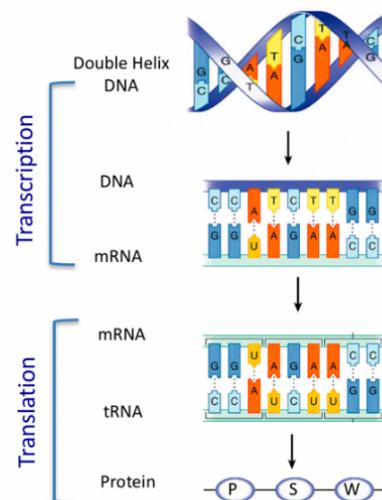
Sentences out of {A,T,C,G}

RNA Language

Sentences out of {A,U,C,G}

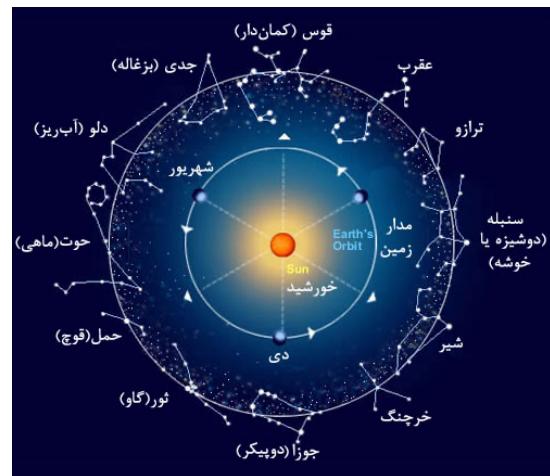
Protein Language

Sentences out of
{A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,V,W,X,Y}



تعريف زبان

Chomsky (1959: 137) “*A language is a collection of sentences of finite length all constructed from a finite alphabet (or, where our concern is limited to syntax, a finite vocabulary) of symbols.*”

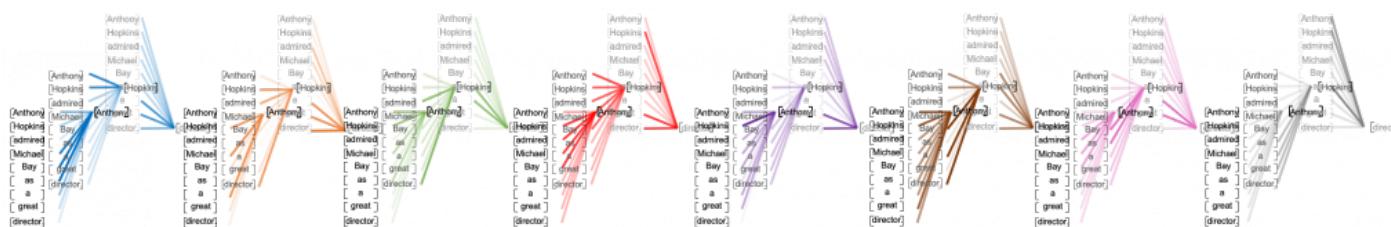
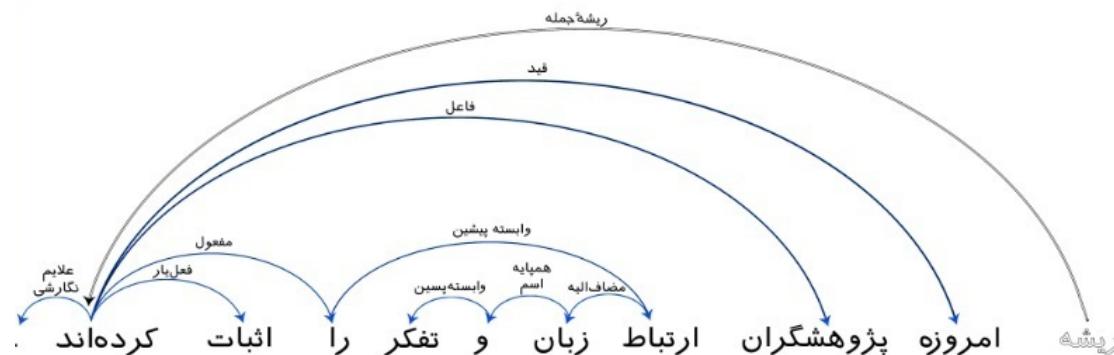


- معرفی
- زبان
- پردازش زبان طبیعی
- NLP درس
- جلسه بعد

تعريف زبان



- مثال ارتباطات درختی و شبکه‌ای



- معرفی زبان
- پردازش زبان طبیعی
- NLP درس
- جلسه بعد

تعریف زبان

Chomsky (1959: 137) “A language is a collection of sentences of **finite length** all constructed from a **finite alphabet** (or, where our concern is limited to syntax, a **finite vocabulary**) of symbols.”



مجموعه توصیفاتی از حقایق: توصیفاتی در محور زمان از المان‌های تکرار پذیر که با هم ارتباطات تنگاتنگ ساختاری و معنایی دارند: ارتباطات درختی و شبکه‌ای.

- معرفی
- زبان
- پردازش زبان طبیعی
- NLP درس
- جلسه بعد

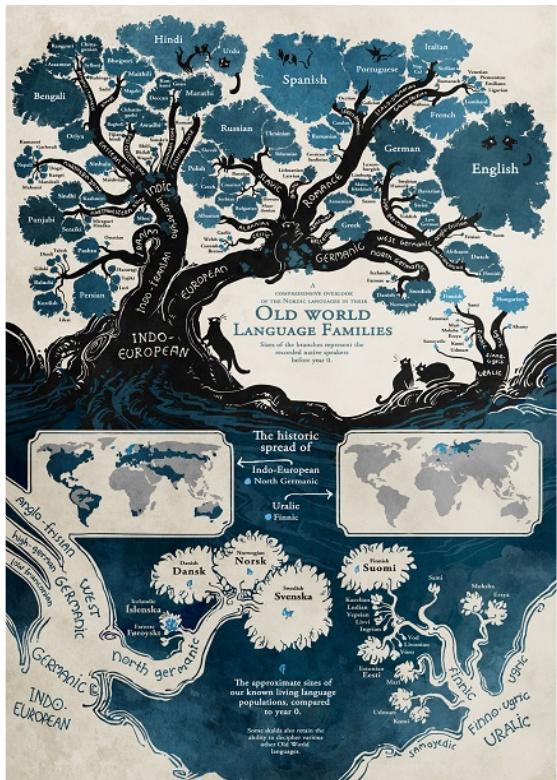
پردازش زبان انسان

◦ حدود ٧٠٠٠ زبان طبیعی

<https://wals.info>

◦ بسیاری با محدودیت منابع

◦ چرا به تکنولوژی زبانی نیاز داریم؟



- معرفی زبان
- پردازش زبان طبیعی
- درس NLP
- جلسه بعد

Languages in the world

There are around 7,000 languages in the world:

- Around 400 languages have more than 1M speakers.
- Around 1,200 languages have more than 100k speakers.
- Africa > 2000 languages & Indonesia 700 languages

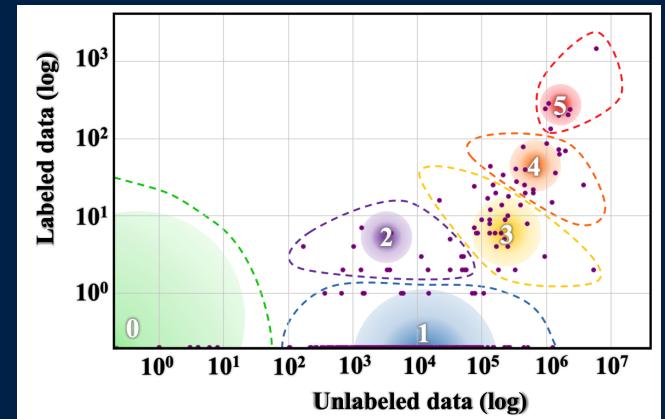


Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara Rivera.
Writing System and Speaker Metadata for 2,800+ Language Varieties.
In Proceedings of the Thirteenth LREC. 2022.



Taxonomy of Languages (i)

- > [LDC catalog](#) and the [ELRA Map](#) for labeled datasets
- > # of [Wikipedia pages](#) for unlabeled data resources



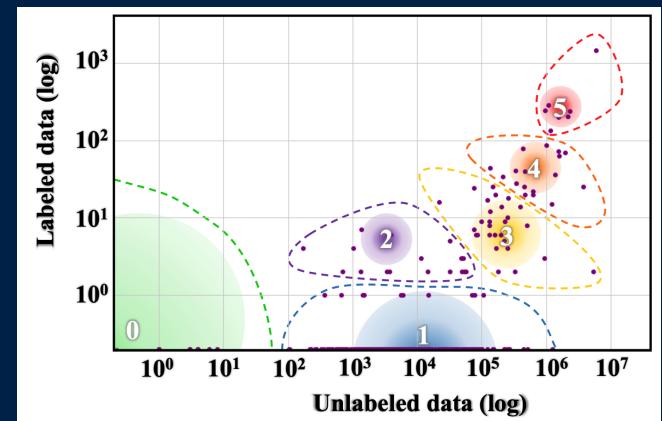
Class	Definition	Example Languages	#Langs	#Speakers	% of Total Langs
0 - The Left-Behinds	Ignored in language tech, limited resources, virtually no unlabeled data, digital upliftment unlikely	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1 - The Scraping-Bys	Some unlabeled data, potential improvement with organized effort, need for awareness and labeled dataset collection	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%



Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury.
[The State and Fate of Linguistic Diversity and Inclusion in the NLP World.](#)
In Proceedings of the 58th ACL, 2020.

Taxonomy of Languages (ii)

- > [LDC catalog](#) and the [ELRA Map](#) for labeled datasets
- > # of [Wikipedia pages](#) for unlabeled data resources



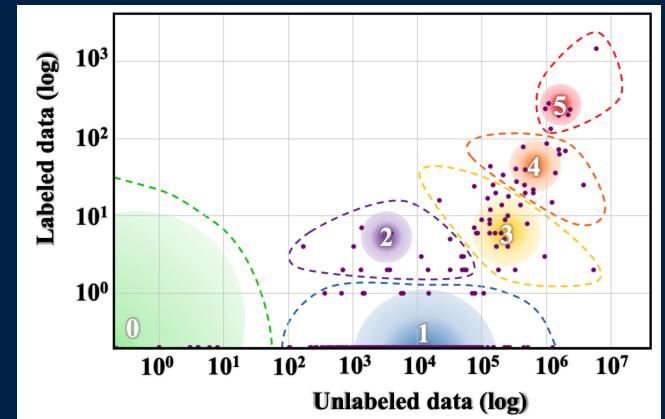
Class	Definition	Example Languages	#Langs	#Speakers	% of Total Langs
2 - The Hopefuls	Small labeled datasets, active research and support communities, promising future with more NLP tools	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3 - The Rising Stars	Benefited from unsupervised pre-training, strong web presence, cultural community online, need for labeled data collection	Indonesian, Ukrainian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%



Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury.
[The State and Fate of Linguistic Diversity and Inclusion in the NLP World.](#)
 In Proceedings of the 58th ACL, 2020.

Taxonomy of Languages (iii)

- > [LDC catalog](#) and the [ELRA Map](#) for labeled datasets
- > # of [Wikipedia pages](#) for unlabeled data resources

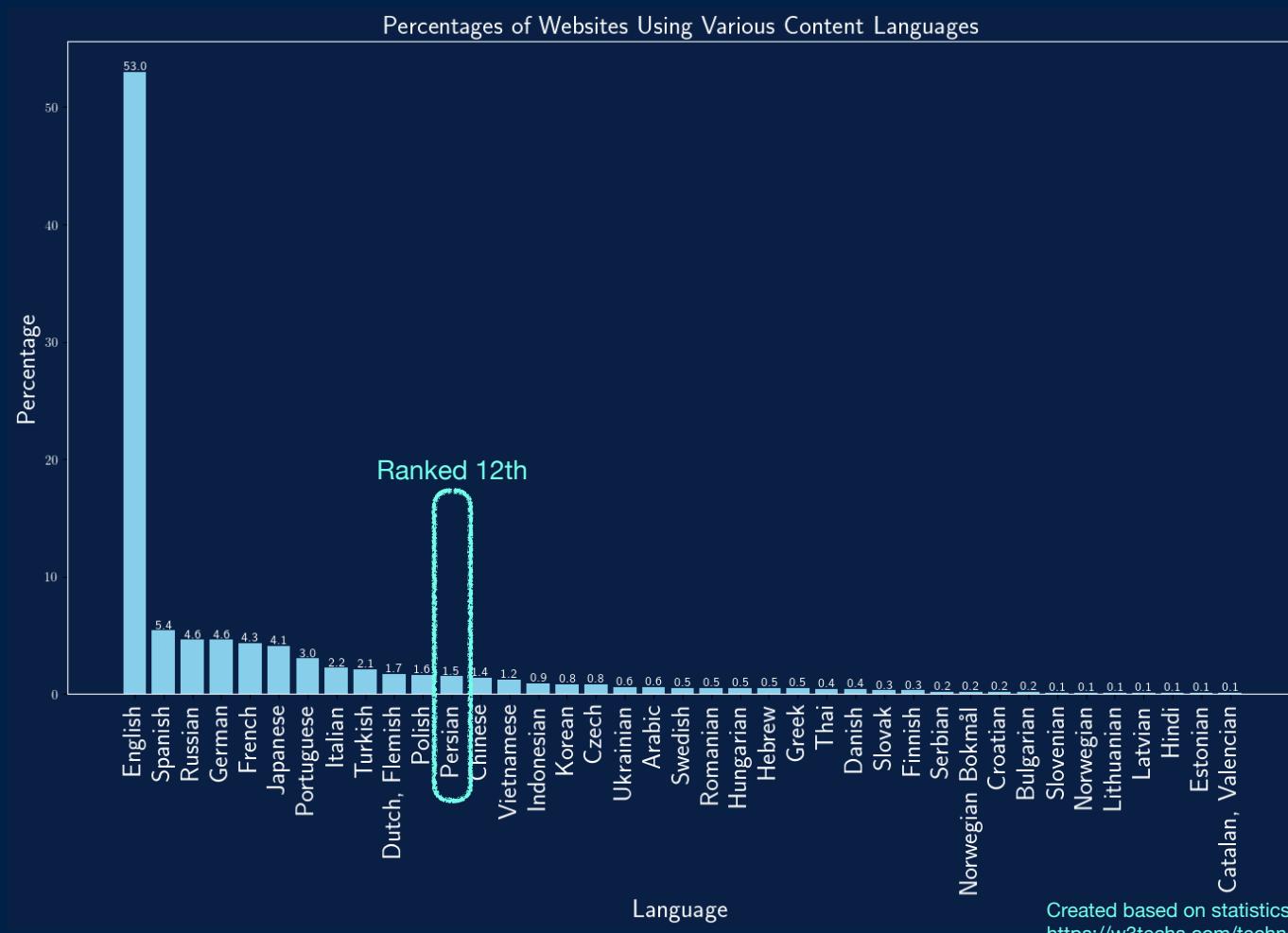


Class	Definition	Example Languages	#Langs	#Speakers	% of Total Langs
4 - The Underdogs	Large unlabeled data, strong resource firepower, active NLP research, potential to reach digital superiority	Persian, Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5 - The Winners	Dominant online presence, extensive industrial and government investment, rich resources and technologies	English, Spanish, German, Japanese, French	7	2.5B	0.28%



Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury.
[The State and Fate of Linguistic Diversity and Inclusion in the NLP World.](#)
 In Proceedings of the 58th ACL, 2020.

Web Content Language Distribution



Created based on statistics provided by w3techs on 3 November 2023
https://w3techs.com/technologies/overview/content_language

Language diversity of Iran

Persian & dialects	58%
Azeri & dialects	26%
Kurdish	9%
Luri	2%
Balochi	1%
Arabic	1%
others	3%



<https://www.internetworldstats.com>



Iranian Languages in Progress @SUT

Marzia Nouri, Mahsa Amani, Reihaneh Zohrabi and Ehsaneddin Asgari

The Language Model, Resources, and Computational Pipelines for the Under-Resourced Iranian Azerbaijani

To be Appeared in ACL 2023.

آذى

Reihaneh Zohrabi, Mostafa Masumi, Omid Ghahroodi, Parham AbedAzad, Hamid Beigy, Mohammad Hossein Rohban and Ehsaneddin Asgari

Borderless Azerbaijani Processing: Linguistic Resources and a Transformer-based Approach for Azerbaijani Transliteration

To be Appeared in ACL 2023.

کردی

Borderless Kurdi Processing

To be submitted.

لری

Luri Language Processing

In progress work.

Multilingual Model for Iranian Languages

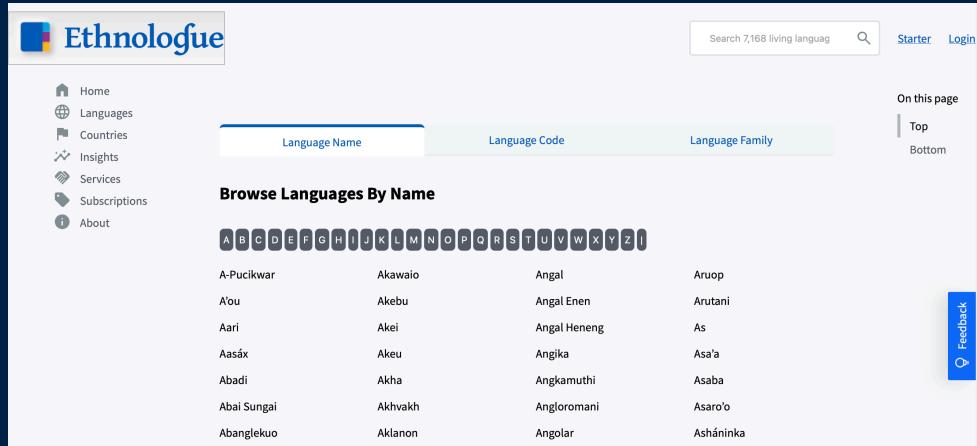
In progress work.





Showing 1 to 100 of 2,662 entries							Previous	1	2	3	4	5	Next	0
Name	WALS code	ISO 639-3	Genus	Family	Macroarea	Latitude	Longitude	Countries						
Search	Search	Search	Search	Search	...any...	Search	Search	Search						
Aari	aar	aiw	South Omotic	Afro-Asiatic	Africa	6.00	36.58	Ethiopia						
Abau	aba	atu	Abau	Sepik	Papunesia	-4.00	141.25	Papua New Guinea						
Abaza	abz	abq	Northwest Caucasian	Northwest Caucasian	Eurasia	44.00	42.00	Russia						
Abenaki (Western)	abw	abe	Algonquian	Algic	North America	44.00	-72.25	Canada United States						
Abidji	abd	abi	Agneby	Niger-Congo	Africa	5.67	-4.58	Côte d'Ivoire						
Abipón	abi	axb	Abipón	Guaicuruan	South America	-29.00	-61.00	Argentina						
Abkhaz	abk	abk	Northwest Caucasian	Northwest Caucasian	Eurasia	43.08	41.00	Georgia						
Abui	abv	abz	Alor-Pantar	Greater West Bornean	Papunesia	-8.25	124.67	Indonesia						
Abun	abu	kgr	Abun	Abun	Papunesia	-0.50	132.50	Indonesia						
Acehnese	ace	ace	Malayo-Sumbawan	Austronesian	Eurasia	5.50	95.50	Indonesia						

<https://wals.info>



<https://www.ethnologue.com>



<https://languages.parsi.ai>



زبانهای ایرانی

سامانه جامع معرفی زبانهای ایرانی و منابع زبانی و زبانشناسی
آزمایشگاه پردازش هوشمند متن و زبان و علوم انسانی محاسباتی



آذری آسما اوستایی بلوچی پارسی تاتی شیرازی طبری کردی گلکنی لری واژی هزدی

بیشترین و کمترین تعداد گویشوران زبان



شمار گویشورهای هر زبان در ایران



نکات جالب در مورد تاریخچه زبان

زبانی که اکنون اغلب مردم آذربایجان به آن تکلم می‌کنند ترکی آذربایجانی نامیده می‌شود. این زبان در نیمهٔ اول قرن پنجم هجری با ترکان سلجوقی وارد قسمت شمال ایران شد. و با کوچ قبیله‌ی غز به آذربایجان و طرح اقتامت و جهانشایی آنان در این میزبان به مروز زمان زبان متداول آذربایجان شد. قدیمی‌ترین نمونهٔ لغت و نظم و نثر عامیانهٔ این زبان در «دیوان لغات الترک کاشغری» و کتاب «داستان‌های دده‌قوقد» به چشم می‌خورد. نخستین نظم عروضی این زبان منسوب به حسین‌اوقل اسفرائین است. با مهاجرت مولانا به آذربایجان و روم شرقی و شیوخ ذذهب تصور در این نقاط ترکی زبان طبقت و وسیله‌ی تبلیغ و ارشاد شد. مولانا جلال الدین می‌گفت:

اگر تاتسان و گر رومسان و گر ترک | زبان بیزبانی را بیاموز
سلطان ول فرزند مولانا مجموعه‌ی غزلیات عارفانه‌ای به زبان ترکی ترتیب داد. نسیمی، فضولی، خطاطی، صائب، قوسی و امانی اشعار و غزلیاتی به این زبان سروdedند. گویندگان بعدی با پیروی از پیشیان بین لغت طبع‌آهانی کردند و در اوائل مشروطه‌ی افکار آزادیخواهانه‌ی غرب در این زبان انگکاس یافت و از راه روزنامه و مجله به مایر نقاط ایران سرایت کرد.

بخش اول بخش دوم بخش سوم

درصد گویشورهای هر زبان در ایران



رشد تعداد گویشوران هر زبان در طول زمان

