



# Advanced and Collaborative AI Research on Arabic

**Mustafa Jarrar**  
Birzeit University  
Palestine

# Artificial Intelligence

Opportunity  
to invest in  
**Arabic**  
as industry  
and  
sustainable  
R&D

The 4<sup>th</sup>  
Revolution

Opportunity  
for  
developing  
countries to  
compete and  
impact



يرجى ضبط المنبه لـ ٩٥٠ جـ ٢  
تشغيل الموسيقى لـ ١٥٠

كيف سيكون الطقس اليوم؟  
أتقل إلى المنزل زل زل زل زل زل  
أرني العواید للأسبوع القبل زل زل زل زل زل

# Huge Market



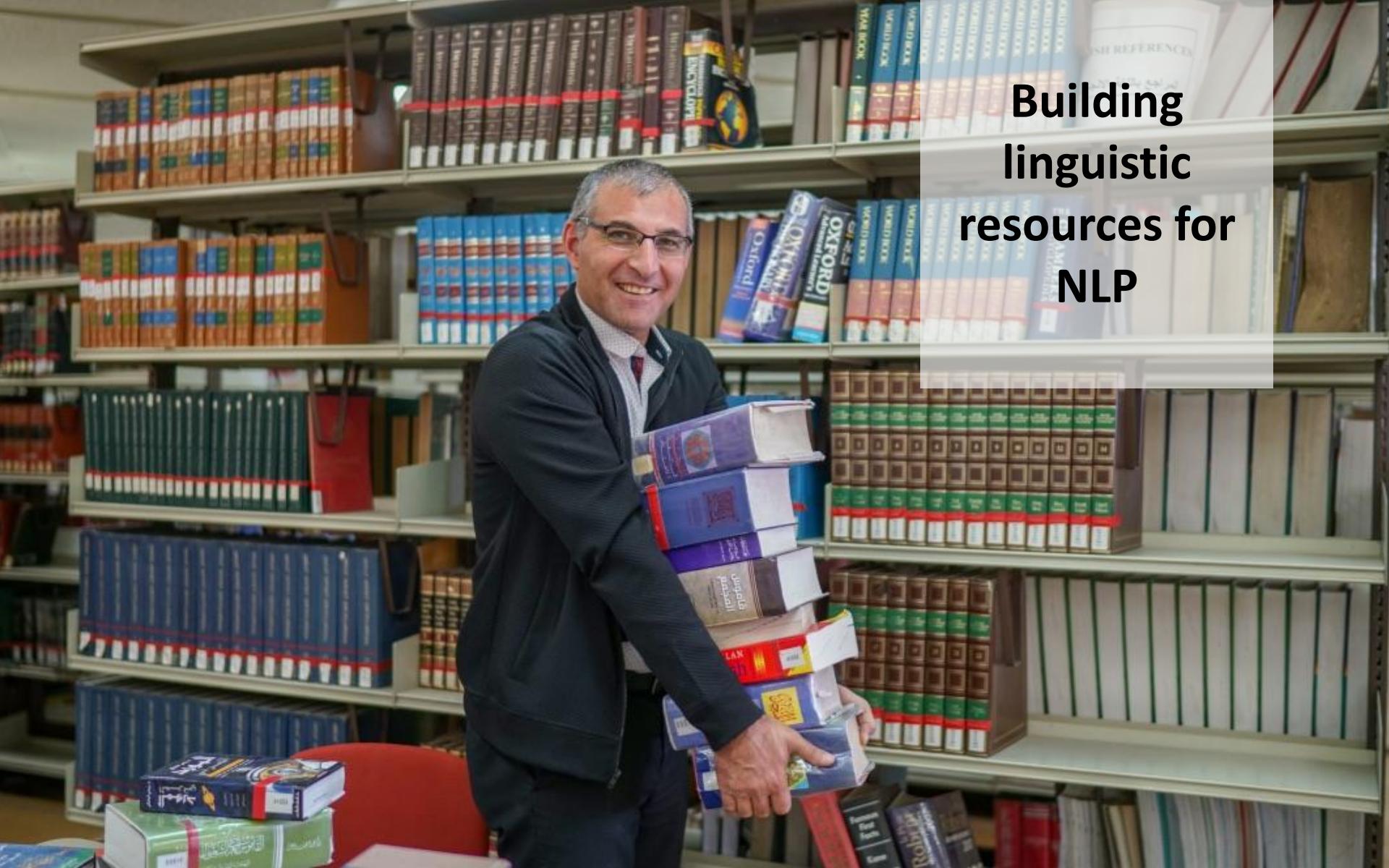
400  
Million  
users  
(as a first  
language)

1.6  
Billion  
users  
(liturgical/  
second  
Language)

Other  
users for  
economic  
and security  
...

# **Consequences of not supporting Arabic in ICT**

**On Education, culture, economy, politics, people with special  
needs, .....**

A photograph of a man with grey hair and glasses, wearing a dark blazer over a white shirt, standing in a library. He is holding a large, tall stack of books against his chest with both hands. The books are of various sizes and colors, including blue, purple, and red. The background consists of several rows of bookshelves filled with books.

**Building  
linguistic  
resources for  
NLP**



# Lexical Resources at Birzeit University

## Lexicographic Database



150 lexicons

Largest Arabic lexicographic database

WSD 84%

## Arabic Ontology/Wordnet



Formal Arabic Wordnet with ontologically clean content

WSD 84%

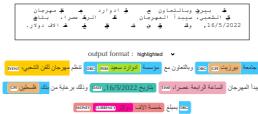
## Annotated Corpora



Arabic Dialects  
Morph. Annotations  
NER , Intents, hate  
....

NER 88.4%

## NLP library



APIs  
Linguistic Data, synonyms, Nested NER, intents, ...

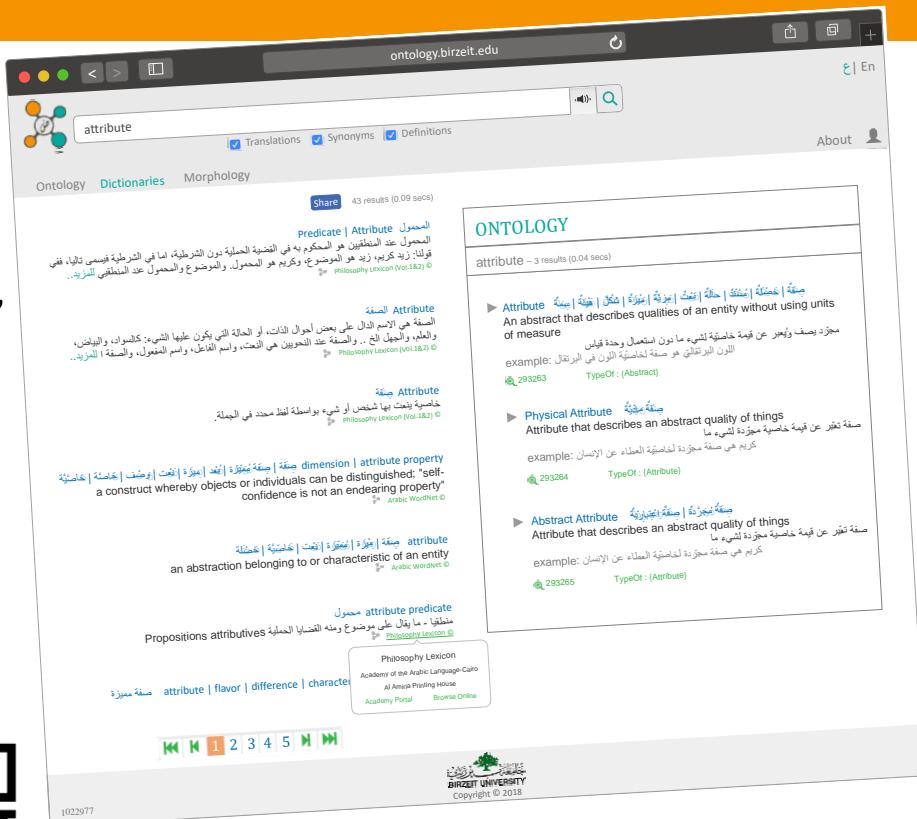


## Big Linguistic Data Graph

<https://ontology.birzeit.edu>

# The Lexicographic Database

- The largest lexicographic Arabic database (150 Lexicons)
- Contains most lexicon types: glossaries, thesauri, bi/trilingual dictionaries, morph datasets, **Arabic Ontology**, and more.
- Covers most domains: science, technology, law, business, art, philosophy, ...
- Digitized over 10 years.



The screenshot shows a web browser window for ontology.birzeit.edu. The search bar contains the word "attribute". Below the search bar are three tabs: "Ontology", "Dictionaries", and "Morphology", with "Dictionaries" currently selected. The main content area displays search results for "attribute".  
**Predicate | Attribute:** This section shows a definition in Arabic: "المحول عند المتطابقين هو المكتوب به في الصيغة الجملية دون الشطرانية، اما في الشطرانية فيسمي ثالثا، ففي قوله زيد كبريه، زيد هو الم موضوع، وكريم هو المحمول، والموضوع والمحمول عند المتطابقين المزدوج".  
**Attribute:** This section shows a definition in Arabic: "الصلة به الاسم الذي على بعض حوار الناس، او الحالة التي يكتبون عليها النصوص: كالساد، والوايفين، والطفل، والجبل، والسماء...، وـالصلة عند المخربين هي الاقتتال، واسم القاتل، واسم المهمول، والصلة (المزدوج)...".  
**Physical Attribute:** This section shows a definition in Arabic: "صفة تغير عن قيمة خاصية مجزأة شئ، مثلاً: ميزان، كرير هي صفة مجزأة لخاصية المعلماء عن الإنسان".  
**Abstract Attribute:** This section shows a definition in Arabic: "صفة تغير عن قيمة خاصية مجزأة شئ، مثلاً: ميزان، كرير هي صفة مجزأة لخاصية المعلماء عن الإنسان".  
At the bottom of the page, there is a footer with the text "Academy of the Arabic Language-Cairo", "Al Amara Printing House", "Academy Portal", "Browse Online", and the date "Copyright © 2018". A QR code is also present at the bottom left.

<https://ontology.birzeit.edu>

# Some Statistics

## 150 Lexicons

Category	Lexical Concepts	Lexical entries	Synsets	Translations pairs	Glosses	Semantic relations
<b>Total (Millions)</b>	1.1 M	2.4 M	1.8 M	1.5 M	0.7 M	0.5 M
<b>Sub Counts</b>		1,100K Arabic 1,100K English 200K French 3K Others 1,300K Single-word 1,000K Multi-word	800K Arabic 800K English 200K French 50K Others	1,000K English-Arabic 300K English-French 200K French-Arabic	400K Arabic 300K English 1K Others	170K Sub-super links 29K Part-of links 260K Has-Domain links 30K Other links

- API accessible for NLP applications.



# Arabic Ontology



- Characterization of the intended meaning (i.e., concepts) that the Arabic terms convey.
- Formal **Arabic Wordnet** - with ontologically-clean content.

- Linked with **PWN**, **Wikidata**, **BFO**, **DOLCE**
- Linked with many lexicons

The screenshot shows a web-based ontology interface. At the top, there is a search bar with a magnifying glass icon and three checkboxes: "Translations", "Synonyms", and "Definitions". Below the search bar are three navigation tabs: "Ontology" (which is active), "Dictionaries", and "Morphology". On the right side of the header, there are links for "About" and "En" (English). The main content area displays a list of entity types with their definitions in English and Arabic, examples, and type information. The entities listed are:

- Entity**: شيء | كيان | كيان Whatever existed or will exist, and can be realized or imagined. أيمالاً موجود أو موجود وشئططيع إدراكه أو تخيله. example: كل شيء على ما يرام. 293198
- Object**: موجود | كيان | أقليم | حقيقي | واقع | شيء | كائن | كيان An entity that is wholly and independently present in time, and is realized either for its concrete or social existence. شيء له ذات مستقلة بنفسه، وحاضر كلّي في الزمن، ويدرك بذلك قياساً أو بذاته اعتباراً يختلف ادراكنا لافي موجود لاختلاف ما يميز أنواعه من المفهومات الجوهريّة. example: 293200 TypeOf : {Entity}
- Occurrent**: تبرورة | حدوث | حصول شيء الذي تدرك ذاته وأجزاءه بجريانه عبر الزمن لا يمكن فهم أي حدث بشكل مفصل عن الأطراف الزمنيّة له. example: 293202 TypeOf : {Entity}
- Dependent Entity**: مُنوط | معمد | مكتفى | مشروط شيء يعتمد وجوده على وجود أشياء أخرى طول المبني منوط بوجود المبني وإلا فلا طول له. example: 293201 TypeOf : {Entity}
- Abstract**: مجردة | مخبرية | غير ملماتي | نظري An entity exists only in mind, cannot be measured or socially realized, and

At the bottom left of the interface, there is a page number "1022977". At the bottom right, there is a logo for "BIRZEIT UNIVERSITY" and the text "Copyright © 2018".

<https://ontology.birzeit.edu/concept/293198>



# Arabic Ontology



- Current size so far (but the numbers are dynamic)

1800 fully-done concepts (mostly top levels)

17K partially investigated (ready for NLP applications)

Some branches are elaborated, other not yet.

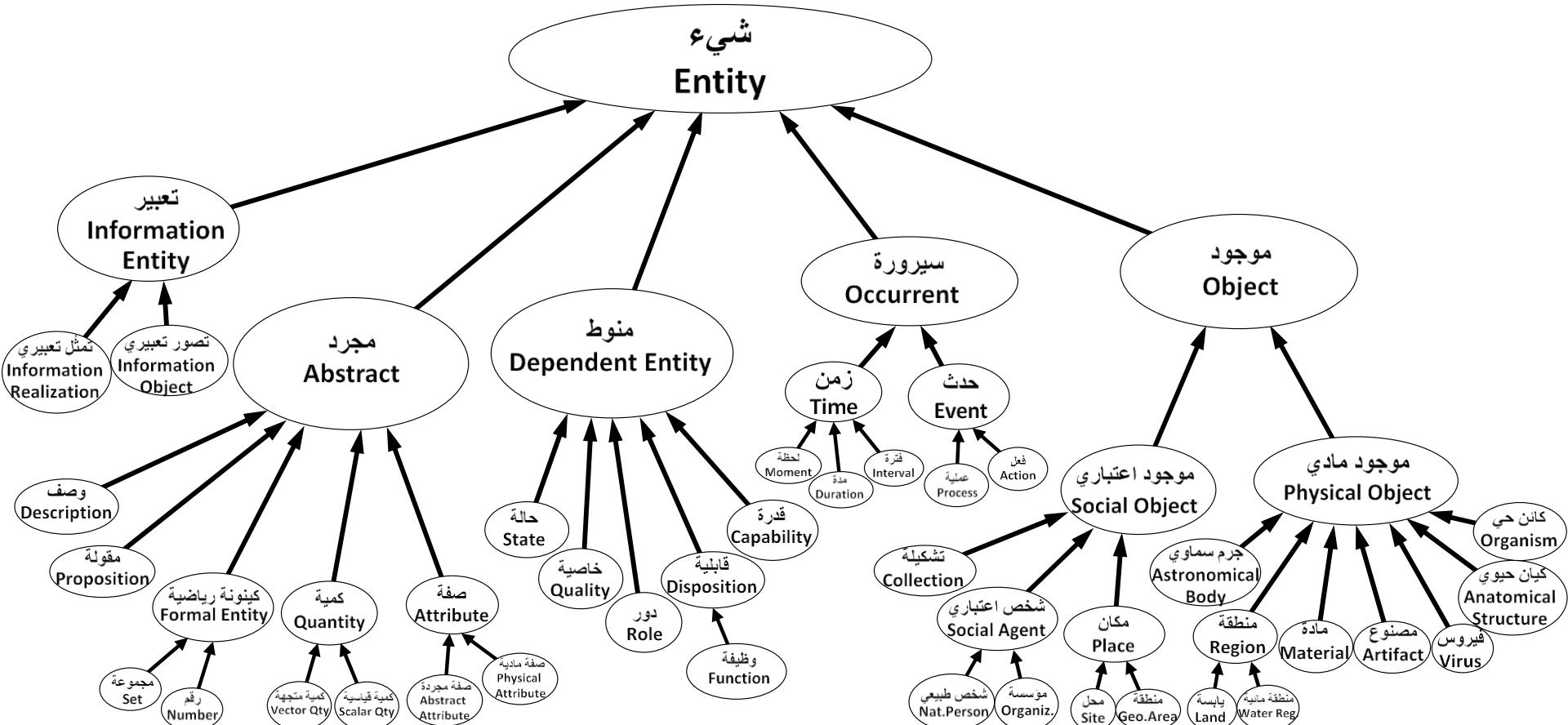
- English labels are not our target - provided for readability and communication.
- Methodology: Built top-down and bottom-up at the same time.

The screenshot displays several concept cards from the Arabic Ontology interface:

- فِيَرْوَرٌ لُّحُوتٌ خَسُولٌ**  
الشيء الذي تدرك ذاته وأجزائه بغير الزمن  
لا يمكن فهم أي حدوث بشكل منفصل عن الإطار الزمني له  
293202 TypeOf : {en ty}
- حدثٌ تراكميٌّ**  
حدث تراكمي يتكون من سلسلة من الأفعال المتتابعة، التي تحدث بشكل متتابع على خط الزمن  
هناك ثلاثة أنواع من الخلايا تساهم في عملية نمو العظمة  
293215 TypeOf : {occurrent}
- عملية بولولوجية**  
عملية بولولوجية  
©Collected  
455000019 TypeOf : {process} Ontologies
- عملية جسمانية**  
عملية جسمانية  
©Collected  
455000574 TypeOf : {biological process} Ontologies
- فُلُوكٌ**  
فُلُوكٌ  
©Collected  
455000020 TypeOf : {bodily process} Ontologies
- عملية إدراكية**  
عملية إدراكية



# Top Levels of the Arabic Ontology



**Curras**, a Palestinian  
morphologically  
annotated corpus (Revised)



**Baladi**, a Lebanese  
morphologically  
annotated corpus



مدونة اللهجات الشامية  
**Levantine Corpus**

65K tokens

El Haff, K., Jarrar, M., Hammouda, T., Zaraket, F., (2022). [\*\*Curras + Baladi: Towards a Levantine Corpus\*\*](#). In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France.

# Annotation Guidelines

بروحي ساكنة ممنون عيونك عيوني وانا عم آخذ المونة رح ظل شوف الدنيا لوما بقيلي عيون DA Sentence:

Token	CODA	Prefix(es)	Stem	Sufix(es)	POS	Person	Aspect	Gender	Number	DA Lamma	MSA Lemma	Gloss
بروحي	بروحي	/ب/PREP	روح/NOUN	/ي/POSS_PRON_1S	NOUN	-	-	m	s	روح	روح 1	spirit
ساكنه	ساكنه		ساكن/ADJ	/ـ/NSUFF_FEM_SG	ADJ	-	-	f	s	ساكن	ساكن 1	residing
ممنون	ممنون		ممنون/ADJ		ADJ	-	-	m	s	ممنون	ممنون 1	grateful
عيونك	عيونك		عيون/NOUN	/ـ/POSS_PRON_2FS	NOUN	-	-	m	1 p	عين	عين 1	eye
عيوني	عيوني		عيون/NOUN	/ـ/POSS_PRON_1S	NOUN	-	-	m	p	عين	عين 1	eye
وانا	وانا	/و/CONJ	ـا/PRON_1S		PRON	1	-	m	s	أنا	أنا 1	I
عم	عم		ـم/PROG_PART		PROG_PART	-	i	-	-	ـم	ـم 1	continue
آخذ	آخذ	/ـ/IV1S	ـخذ/IV		VERB	1	i	m	s	ـأخذ	ـأخذ 1	take
المونة	المونة	/ـ/DET	ـمون/NOUN	/ـ/NSUFF_FEM_SG	NOUN	-	-	f	s	ـمـونة	ـمـونة 1	provisions
رح	رح		ـرح/FUT_PART		FUT_PART	-	-	-	-	ـرح	ـرح 1	will
ظل	ظل		ـظل/IV		VERB	1	i	m	s	ـظل	ـظل 1	remain
سوف	سوف		ـسوف/IV		VERB	1	i	m	s	ـشـاف	ـشـاف 1	looking
الدنيـا	الدنيـا	/ـ/DET	ـالـدنيـا/NOUN		NOUN	-	-	f	s	ـدـنـيـا	ـدـنـيـا 2	world
لومـا	لومـا		ـلومـا/CONJ		CONJ	-	-	-	-	ـلـوـما	ـلـوـما 1	except for
بقـيلي	بقـيلي		ـبـقـيلي/IV	/ـي/IVSUFF_SUBJ:1S	VERB	3	i	m	s	ـبـقـيـ	ـبـقـيـ 1	remain
عيـون	عيـون		ـعيـون/NOUN		NOUN	-	-	m	p	ـعـيـنـ	ـعـيـنـ 1	eye

# Levantine Corpus



هيك										Search
<input checked="" type="checkbox"/> Word <input checked="" type="checkbox"/> Stem <input checked="" type="checkbox"/> MSA <input checked="" type="checkbox"/> Surface <input checked="" type="checkbox"/> Gloss <input type="radio"/> Whole Word <input checked="" type="radio"/> Substring <input checked="" type="checkbox"/> Palestinian <input checked="" type="checkbox"/> Lebanese										
										سایت ... بمحض اسراره ون میں من سام رہ۔ یو... ... معروف فس نفس . بتطلع الخميس ليش هالقدة الازمة بقولك يوم الخميس دايما هيك ، هذا يوم في الاسبوع طاير ...
هيك	هيك	هيك	هيك 0	هكذا 1	هيك/اداة استفهام + ظرف	أداة استفهام + ظرف	لا ينطق	لا ينطق	"is + this way,thus   this way,this,thus"	الجار : يا زلمة دوشتنا طرشتنا ياعن ابو هيك سكتة
هيك	هيك	هيك	هيك 0	هكذا 1	هيك/اداة استفهام + ظرف	أداة استفهام + ظرف	لا ينطق	لا ينطق	"is + this way,thus   this way,this,thus"	يلم اللي اكلناه ، عزمنا الاسبوع الماضي ، عشان نردهم العزيمة ، وهيك بذك تحطيلهم ترترسي
وهيك	وهيك	هيك?	هيك 1-	هكذا 1	و/عطف + هيك/اسم موصول	اسم موصول	ذكر	لا ينطق	so; thus; such	رجل 2 : هيك مزبطة بدها هيك ختم
هيك	هيك	هيك	هيك 0	هكذا 1	هيك/اداة استفهام + ظرف	أداة استفهام + ظرف	لا ينطق	لا ينطق	"is + this way,thus   this way,this,thus"	رجل 2 : هيك مزبطة بدها هيك ختم
هيك	هيك	هيك	هيك 0	هكذا 1	هيك/اداة استفهام + ظرف	أداة استفهام + ظرف	لا ينطق	لا ينطق	"is + this way,thus   this way,this,thus"	معروف فس نفس . بتطلع الخميس ليش هالقدة الازمة بقولك يوم الخميس دايما هيك ، هذا يوم في الاسبوع طاير ....
هيك	هيك	هيك	هيك 0	هكذا 1	هيك/اداة استفهام + ظرف	أداة استفهام + ظرف	لا ينطق	لا ينطق	"is + this way,thus   this way,this,thus"	... معروف فس نفس . بتطلع الخميس ليش هالقدة الازمة بقولك يوم الخميس دايما هيك ، هذا يوم في الاسبوع طاير ....
هيك	هيك	هيك	هيك 0	هكذا 1	هيك/اداة استفهام + ظرف	أداة استفهام + ظرف	لا ينطق	لا ينطق	is + this way,thus   this way,this,thus	... معمول في ناس هيك ما بتکفر بعضها ! عالم فاضية عفكرة بال... عنجد ،
هيك	هيك	هيك	هيك 0	هكذا 1	هيك/اداة استفهام + ظرف	أداة استفهام + ظرف	لا ينطق	لا ينطق	is + this way,thus   this way,this,thus	صیر تدلني امرار ع اطياق ما فيهن لحمة عشان انو ما كون محدود بخياراتي و هيك .
هيك	هيك	هيك	هيك 0	هكذا 1	هيك/اداة استفهام + ظرف	أداة استفهام + ظرف	لا ينطق	لا ينطق	is + this way,thus   this way,this,thus	... هاش المكونات ، قالتلي انتبه هيدي يمكن فيها كحول و أنت يمكن ما بتناكل هيك شي ، و اختارتني وحدة ما فيها كحول ....
هيك	هيك	هيك	هيك 0	هكذا 1	هيك/اداة استفهام + ظرف	أداة استفهام + ظرف	لا ينطق	لا ينطق	is + this way,thus   this way,this,thus	... ي ، ان مش ساعدة بيع بانج انت ... بس في شي تغیر ، شي مش عارف شو هوی ، هيك شي اختفى مثل لما اكل اكلة مش بالبيت وحس...

<https://portal.sina.birzeit.edu/curras>

# Wojood

## Nested Arabic Named Entity Corpus

Flat NER

Sami	works	at	Jimmy	Carter	Center
B-PERS	O	O	B-ORG	I-ORG	I-ORG

✓mature

Nested NER

Sami	works	at	Jimmy	Carter	Center
B-PERS	O	O	B-PERS	I-PERS	I-ORG

!

Jarrar, M., Khalilia, M., Ghanem, S. (2022). [Wojood: Nested Arabic Named Entity Corpus and Recognition using BERT](#). In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France.

# Wojood

## ❖ NER Corpus

Corpus	Nested?	size (tokens)	No. of entities	Entity classes	Which Arabic	Domain
Ontonotes5	No	300k	28k	18	MSA	News
ANERCorp	No	150k	11k	4	MSA	News
Canercorpus	No	258k	72k	14	Classic	Religion
AQMAR	No	74k	-	open	MSA	4 domains
Wojood Corpus	YES	550K	75K	21	MSA & Dialect	Multi domains Media, History, Culture, Health, Finance, ICT, Law, Elections, Politics, Migration, Terrorism

## ❖ NER BERT

- Multi-task learning (nested entities)
- 88.4% F1-score

# Annotation Guidelines

## 21 entity classes

PERS

People names

فيروز، عادل إمام، ابن احمد، الملك عبدالله، النبي محمد

NORP

Group of people

العرب، المسيحيين، سكان القدس، وزراء الخارجية العرب

OCC

Occupation or professional title

رئيس جامعة ييرزيت، مدير بنك فلسطين، قائد الجيش

ORG

Legal/social body

بنك القاهرة، ريال مدريد، داعش، الجيش المصري،

GPE

Geopolitical: country, city, state

ليبيا، مدينة القدس، الجمهورية اللبنانية، روسيا الاتحادية

LOC

Geographical location (non-GPE)

البحر الميت، قناة السويس، آسيا، الوطن العربي

FAC

Places: landmark, road, building..

مطار صنعاء، سجن ابو غريب، المسجد الأقصى

EVENT

Events of general interest

حرب 1973، القمة العربية 2005، عيد الفطر ، يوم الأرض

DATE

Specific/relative date (>day)

1 يونيو، 2019-2020، الفترة العثمانية

TIME

Specific/relative time (<day)

الساعة ١٢، من الساعة ٥ حتى ٧ مساء، خلال ساعتين

LANGUAGE

Human language or dialect

اللغة العربية، الفصحي، الدارجة المغربية، اللغة الفرنسية

WEBSITE

Website or specific URL

موقع فيسبوك، يوتوب، schema.org

LAW

Geographical location (non-GPE)

قانون الاستثمار ، المادة 114 من قانون العقوبات 2005

PRODUCT

Vehicle, weapon, food, ...

مرسيدس سي ١٨٠، ايفون ١٣ ، دبابه مركابا، تروفين

CARDINAL

Numerals in digits/words

١٥٠، صفر، اثنان ، أربعة وعشرون ، مليون

ORDINAL

does not refer to a quantity

٣ كيلومتر، مئة قدم ، ٣ طن ، ٥٥ غرام ، ٢٥ سم مربع

PERCENT

Word/symbol refers to a percent

٥ بالمئة ، ١٠٪ ، ٩ من كل الف

QUANTITY

Value measured by units

٣ كيلومتر، مئة قدم ، ٣ طن ، ٥٥ غرام ، ٢٥ سم مربع

UNIT

Name/symbol of a unit

ميل ، كيلو ، كيلومتر ،إنش ، كيلوغرام ، هكتار ، مل

MONEY

Monetary quantity, incl.

مئة وخمسون درهم اماراتي ، اثنان وثلاثون يورو ، ٨ دولار

CURR

Name/symbol of currency

دولار ، جنية مصرى ، دينار ، فرنك ، ريال سعودي € ،

# Corpus Collection

Source - Topics	Sentences	Tokens
Web Articles <sup>1</sup> (MSA) Health, Finance, ICT, Law, Elections, Politics, Migration and Terrorism	9,053	258,102
Archive <sup>2</sup> (MSA) History and Culture	12,271	227,020
Social Media <sup>3</sup> (Dialect) General topics	5,653	65,342
<b>Total</b>	26,977	550,464

<sup>1</sup> un.org, hrw.org, msf.org, who.org, mipa.institute, elections.ps, sa.usembassy.gov, diplo- matie.ma, quora.com ....

<sup>2</sup> Awraq, Birzeit University Digital Palestinian Archive

<sup>3</sup> Palestinian and Lebanese dialect corpora

# Nested NER Results

Tag	Precision	Recall	F1-Score
PERS	0.9135	0.9122	0.9129
NORP	0.6828	0.7037	0.6931
OCC	0.7993	0.8402	0.8193
ORG	0.8924	0.9072	0.8997
GPE	0.9424	0.9516	0.9470
LOC	0.8054	0.7059	0.7524
FAC	0.7366	0.6481	0.6895
PRODUCT	0.3333	0.2500	0.2857
EVENT	0.6364	0.6488	0.6425
DATE	0.9253	0.9394	0.9323
TIME	0.6000	0.5122	0.5526
LANGUAGE	0.9310	0.7105	0.8060
WEBSITE	0.4496	0.5472	0.4936
LAW	0.8525	0.9123	0.8814
CARDINAL	0.8437	0.8575	0.8505
ORDINAL	0.9411	0.9448	0.9430
PERCENT	0.2903	0.9310	0.4426
QUANTITY	0.2500	0.1667	0.2000
UNIT	0.5000	0.1667	0.2500
MONEY	0.9143	0.8205	0.8649
CURR	0.8810	0.9487	0.9136
Overall	0.8772	0.8909	0.8840

Entities of the same time  
are not supported yet

# Downloads and Demo



<https://ontology.birzeit.edu/wojood>

بـ بيرزيت وبالتعاون مع ادوارد سعيد مهرجان  
الشعبي، سيبدأ المهرجان في الرابعة عصراً، بتاريخ  
16/5/2022، وذلك بـ بنك فلسطين بـ ٥ الاف دولار.

output format : highlighted ▾

جامعة بيرزيت GPE ORG وبالتعاون مع مؤسسة PERS ادوارد سعيد EVENT تنظم مهرجان للفن الشعبي،

سيبدأ المهرجان الساعة الرابعة عصراً، بتاريخ DATE 16/5/2022، وذلك برعاية من TIME بنك فلسطين GPE

بمبلغ MONEY خمسة الاف CURRENCY دولار ORG

# ArabGlossBERT

## Arabic Word-Sense Disambiguation

### The Word Sense Disambiguation (WSD) Task

Given a word in a context, which sense (i.e. meaning) this word denotes?

قصيدة من عيون الشعر

#### Set of senses

1. عُضو الإبصار في الإنسان والحيوان: له عينان كعَيْنَيِّ الصقر - لا إنما العينان للقلب رائِدٌ ...
2. جاسوس، "كان عيناً لدولة أجنبية". بِتَّ العيون : تجسس، راقب - فلان عَيْنٌ على فلان : ناظر عليه
3. أجود كل شيء وأحسنها ونقيسها: عيون الفن.
4. حارس: فلان عين على المكان.
5. الحاضر من كل شيء أصبح أثراً بعد عين ...
6. عَيْنُ الماء:- ينبوءه، تُحلق الطيور فوق عيون الماء
7. عَيْنُ الشَّيْء:- نفسه، ذاته ( تستعمل للتوكيد): جاء القوم أعينهم - كُنّا في المكان عينه.
8. عَيْنُ العقل:- قدرة ذهنية موروثة على التخيّل وتذكر الأحداث.
- ..... .9

# ArabGlossBERT

## Arabic Word-Sense Disambiguation

### ❖ Arabic context-gloss pairs Dataset (167k)

- Extracted from Birzeit University's Lexicographic database
- Annotated target words in context;

Gloss	Context	Label
[CLS] قصيدة من عيون الشعر [SEP]	أجود كل شيء وأحسنه ونفيسيه [SEP]	True
[CLS] قصيدة من عيون الشعر [SEP]	عين الشيء : نفسه ، ذاته ( تستعمل للتوكيد ) [SEP]	False
[CLS] جاء القوم أعينهم [SEP]	عين الشيء : نفسه ، ذاته ( تستعمل للتوكيد ) [SEP]	True
[CLS] جاء القوم أعينهم [SEP]	أجود كل شيء وأحسنه ونفيسيه [SEP]	False

### ❖ Three Fine-tuned BERT Models

- WSD into **binary sequence-pair classification task**
- **Accuracy 84%**
- 4 types of signals to emphasize target words in context

# Fine-Tuning

- ❖ Fine-tuned three Arabic pre-trained BERT models  
The WSD task is converted into **binary sequence-pair classification task**

[CLS] قصيدة من عيون الشعر [SEP]	أجود كل شيء وأحسنـه ونفيـه	[SEP]	True
[CLS] قصيدة من عيون الشعر [SEP]	عين الشيء : نفسه ، ذاته ( تستعمل للتوكيد )	[SEP]	False
[CLS] جاء القوم أعينهم [SEP]	عين الشيء : نفسه ، ذاته ( تستعمل للتوكيد )	[SEP]	True
[CLS] جاء القوم أعينهم [SEP]	أجود كل شيء وأحسنـه ونفيـه	[SEP]	False

- ❖ Results

Model	Precision	True	False	Accuracy
AraBERTv02	Precision	81	85	84
	Recall	66	93	
	F1-score	72	89	
CAMELBERT	Precision	77	83	82
	Recall	60	92	
	F1-score	67	87	
QARiB	Precision	73	82	80
	Recall	58	90	
	F1-score	65	86	

# Constructing a dataset of context-gloss pairs

## Statistics:

	<b>count</b>
Unique Lemmas (undiacritized)	26169
Avg glosses per Lemmas	1.25
Unique Glosses	32839
Unique Contexts	60272
Avg context per gloss	1.83
True context-gloss pairs	60323
False context-gloss pairs	106884
Total True and False pairs	167207

## Training and Test Datasets

- every context selected in the test set should not be selected in the training set;
- every gloss should be selected in both the training and the test sets.

Datasets	Pairs	Count	Total
Training	True pairs	55,585	152,035
	False pairs	96,450	
Test	True pairs	4,738	15,172
	False pairs	10,434	
	<b>Total</b>	167,207	

Download: <https://ontology.birzeit.edu/downloads>

# ArabGlossBERT

## Arabic Word-Sense Disambiguation



Arabic Language Understanding Tools

..... Confirm

قصيدة من عيون الشعر

Tokenize Lemmatize NER WSD

• قصيدة [قصيدة 1 \(303044571\)](#): مجموعة من الأبيات الشعرية متّحدة في الوزن والقافية والرّوّي وهي تتكون من سبعة أبيات فاكثر "قصيدة غزليّة".  
بيّت القصيدة : البيت المتضمّن غاية الشّاعر، أو أنفس أبياتها، أو مثل يُصرّب في تفضيل بعض الشّيء على كُلّه - مطلع القصيدة : أول بيت منها

• من (من 1)

• عيون [عيون \(غين 2\) \(303038477\)](#): أجود كلّ شيء وأحسنه ونفيسه "قصيدة من عيون الشعر - عيون الفنّ".

• الشعر [الشعر \(ثيغر 1\) \(303029103\)](#): كلام موزون مقوّى قصدًا يعتمد على التخييل والتّأثير؛ ليوحى بإحساسات مؤثرة وصور خيالية "شعر صافي الديباجة - نظم الشّعر - ما الشّعر إلّا شعور الماء يُرسّله ... عفو الديباجة عن صدق وإيمان - إنَّ من الشّعر لِحكمة [ الحديث ] - حِمَّا عَلَّمَنَا الشّعرَ وَمَا يَنْبَغِي لَهُ يَسِّرٌ 69 ". أنشده الشّاعر : قرأه عليه - أوايد الشّاعر : ما لا تماثل جودته أو قوافيه الشاردة - ربة الشّاعر : إلهة الشّعر عند الوثنين - شطرا بيت الشّاعر : الصدر

https://ontology.birzeit.edu/tools/ALMA.html

BIRZEIT UNIVERSITY

جميع الحقوق محفوظة © 2022 جامعة بيرزيت

27




Search Arabic and English terms ...

Translations
 Synonyms
 Definitions

Ontology
Dictionaries
Morphology

About
License



# Visit Our NLP Library



<https://ontology.birzeit.edu/lexapi>

## NLP Library

Download and access NLP data, corpora, tools and services

We developed hundreds of RESTful web services (APIs) for other third-party software developers to directly download and access our lexicographic databases, corpora, and tools. Some are listed on this page. Please get in touch with us for details.

The “sampleKey” API token allows you to use all web services but with a limited number of requests (enough to try it!), if you need more requests, you need to [Request API Token](#).

### APIs:

[\*\*+ Search 150 dictionaries \(synonyms, translations, glosses\)\*\*](#)

[\*\*+ Arabic Ontology\*\*](#)

[\*\*+ Dialect Corpora\*\*](#)

[\*\*+ Word Sense Disambiguation\*\*](#)

[\*\*+ Arabic Named Entity Recognition \(Woood\)\*\*](#)

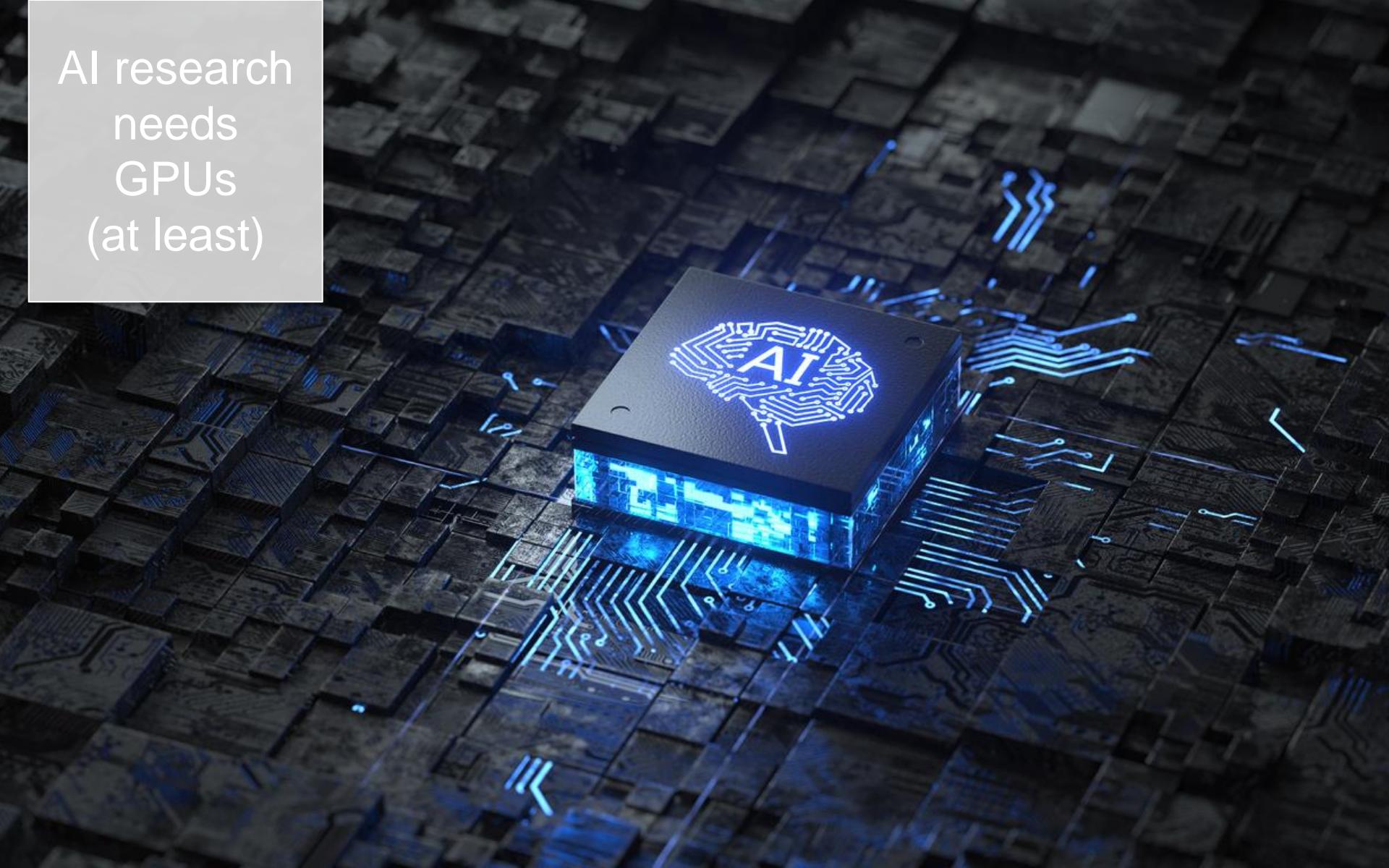
[\*\*+ Lemmatizer\*\*](#)

[\*\*+ Retrieve Morphology Information \(MSA and Dialects\)\*\*](#)

[\*\*+ Arabic Processing Tools\*\*](#)

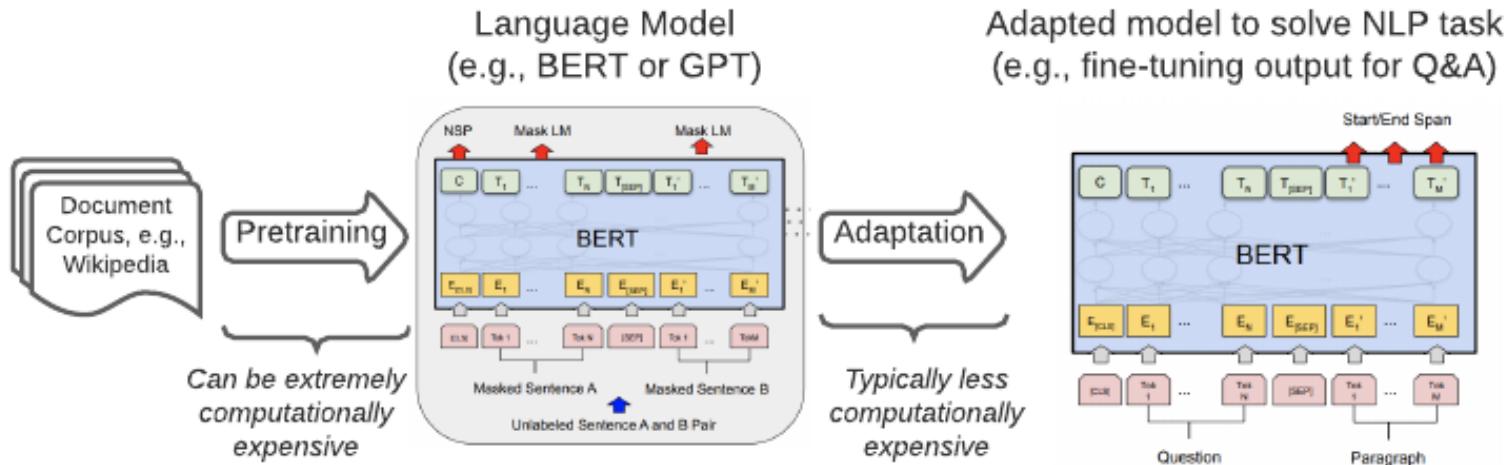
[\*\*+ Autocomplete Service \(Arabic and English\)\*\*](#)

AI research  
needs  
GPUs  
(at least)



# Computing Resources

## Deep Learning



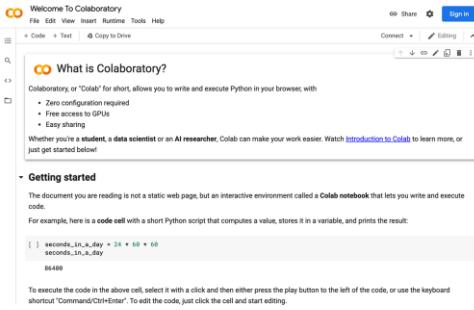
AI research Language is not possible without **fast GPUs**

Each experiment takes **many, many hours/days**

# Cloud is expensive of students



## Google Cloud



A screenshot of the Google Colaboratory interface. At the top, it says "Welcome To Colaboratory". Below that is a navigation bar with "File", "Edit", "View", "Insert", "Runtime", "Tools", and "Help". A "Sign in" button is also present. The main area has tabs for "Code" and "Text", with "Code" currently selected. It shows a snippet of Python code: `seconds\_in\_a\_day = 24 \* 60 \* 60` followed by the output `86400`. Below the code cell, there's a note about Colaboratory being an interactive environment called a "Colab notebook". It lists benefits like zero configuration required, free access to GPUs, and easy sharing. It also encourages users to watch an introduction video. At the bottom, there's a "Getting started" section with a note that the document is an interactive environment and a code cell showing a simple calculation.

Currently, each student have to  
**pay 50\$/month to Google**

Google also offer Colab

# GPUs are not good enough!

CPU



GPU



TPU

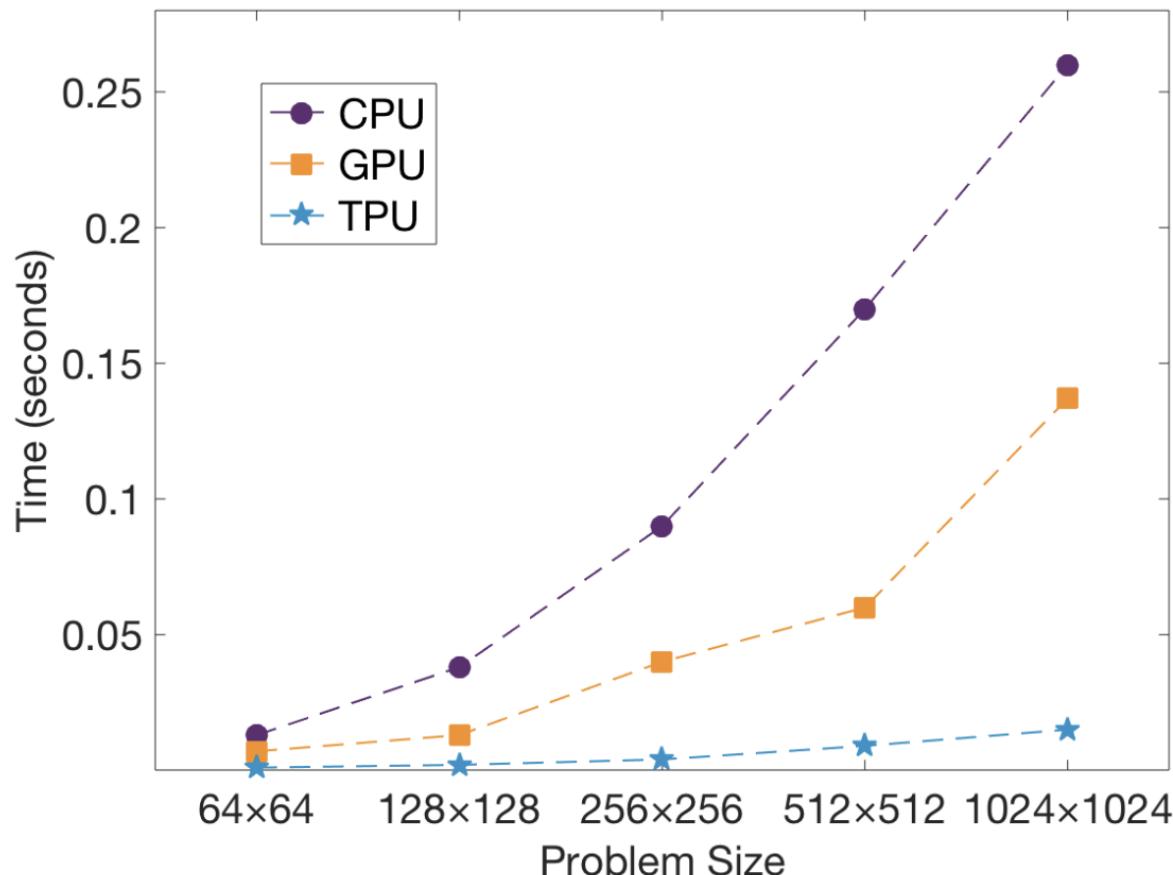


Takes hours/days  
to train a language  
model



Really fast!

# CPU vs GPU vs TPU



# Recommendations

**Build own TPU Cloud**

**For top 20 AI centers  
in the Arab countries**

**Negotiate Google/Amazon**

**For affordable prices  
for more researchers**

# Thank You

Mustafa Jarrar  
[mjarrar@birzeit.edu](mailto:mjarrar@birzeit.edu)