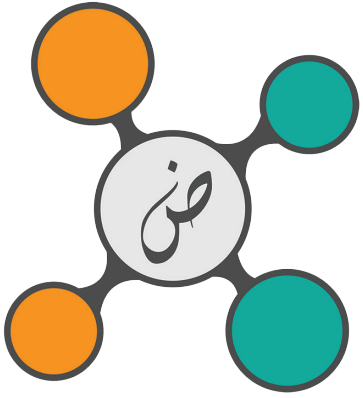


Arabic Natural Language Understanding as a Sustainable AI Research

Mustafa Jarrar
Birzeit University, Palestine

Artificial Intelligence

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

The 4th Revolution

Opportunity for developing countries to compete and impact


يرجى ضبط المنبه
تشغيل الموسيقى
كيف سيكون الطقس اليوم
انتقل إلى المنزل
أرنى المواعيد للأسبوع المقبل

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, ……

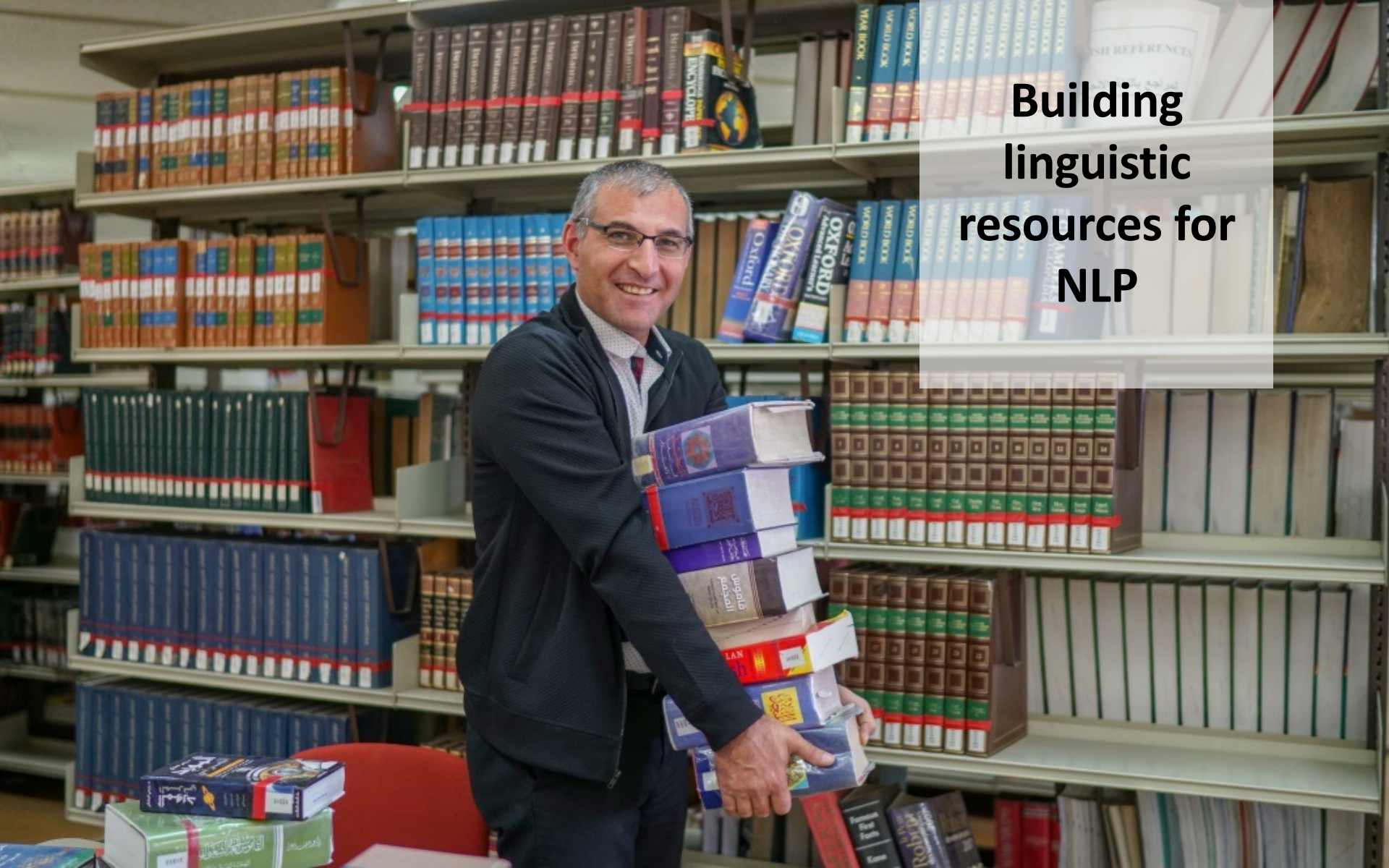
Why Lexical Resources!

- The importance of lexical resources (dictionaries, thesauri, wordnets, linguistic ontologies) is increasing in many application areas, such as:
  - NLP tasks and applications
  - Information search and retrieval
  - Multilingual big data
  - Multilingual semantic web
  - Data integration
  - among many others.
- Lack of Arabic Lexical resources for human use!
- Lack of Arabic Lexical resources for NLP!

At Birzeit University

Digitize, Collect, Build, then clean and link

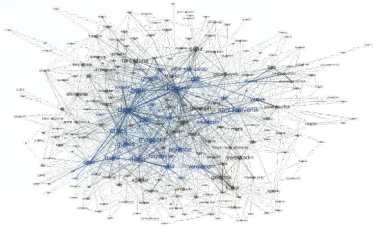
- Make available online for people
- Make available through APIs for NLP applications


Building linguistic resources for NLP

جامعة بيرزيت
BIRZEIT UNIVERSITY

Lexical Resources at Birzeit University

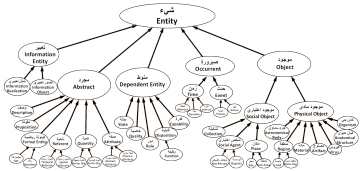
Lexicographic Database



150 lexicons

Very large Arabic-multilingual database

Arabic Ontology



Formal Arabic Wordnet

with ontologically clean content

Dialect Corpora

Annotated corpora

each word is annotated with many morph features

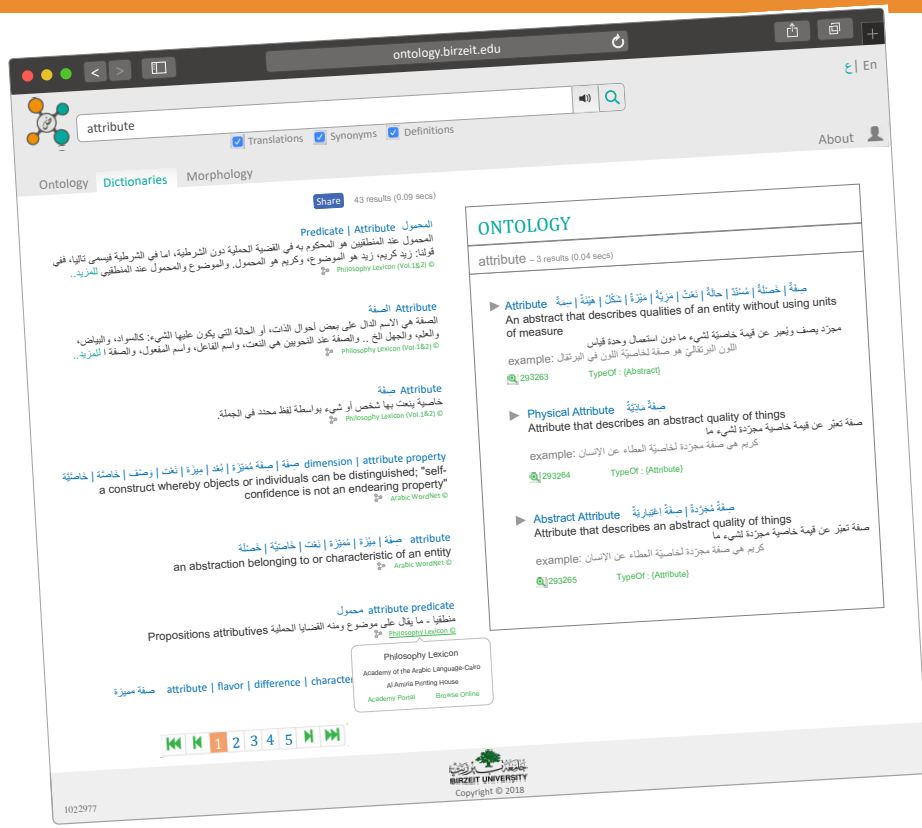
Big Linguistic Data Graph

https://ontology.birzeit.edu

Lexicographic Search Engine

The Lexicographic Database

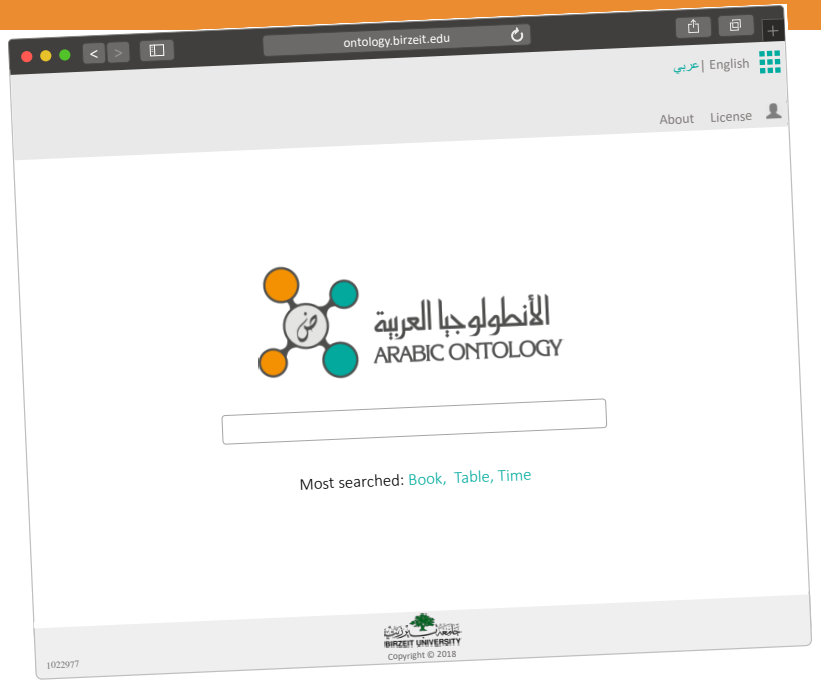
- **The largest lexicographic Arabic database**
- **Contains most lexicon types**: glossaries, thesauri, bi/trilingual dictionaries, morph datasets, **Arabic Ontology**, and more.
- **Covers most domains:** science, technology, law, business, art, philosophy, ...



https://ontology.birzeit.edu

Lexicographic Search Engine

- **Free access to people**: students, translators, researchers, Arabic learners …
- **API accessible** for NLP applications.
- **Done over 9 years (No funding!)**



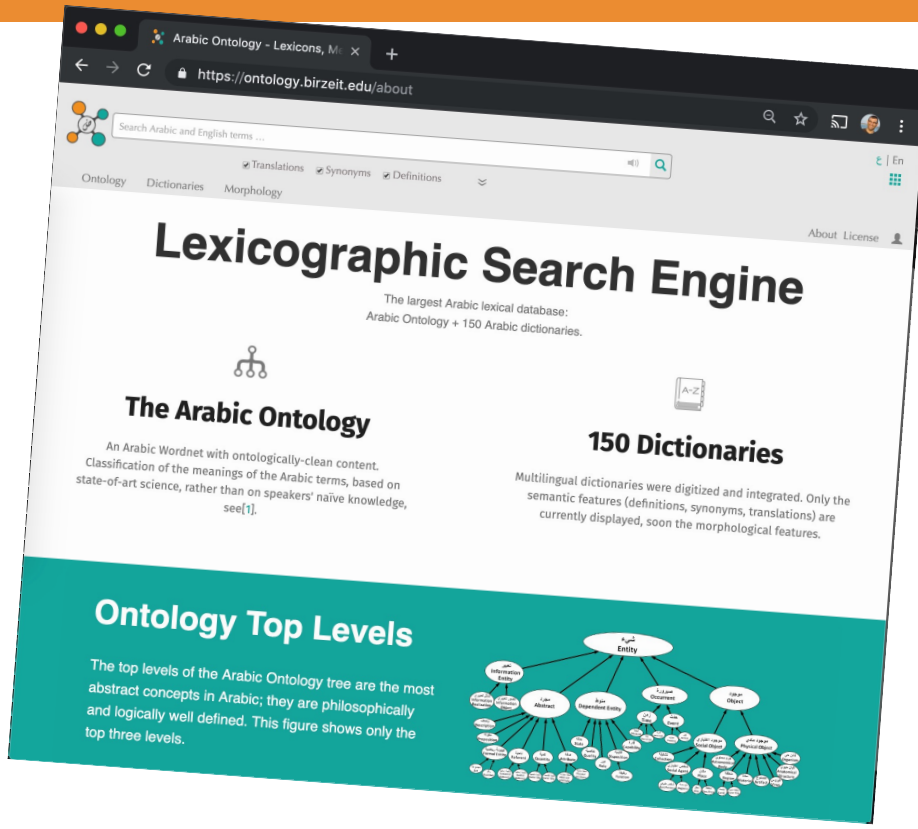
https://ontology.birzeit.edu

Reference:

Mustafa Jarrar, Hamzeh Amayreh: **An Arabic-Multilingual Database with a Lexicographic Search Engine.** NLDB 2019. Pages(234--246), LNCS 11608, Springer. 2019.

Lexicographic Search Engine

- **Search 150 lexicons** for definitions, synonyms, specialized translations, morphology, ontology [3,4] ...
- **Accurate**! compared with machine translation.
- **The first of its kind**! e.g., there are no similar search engines for English lexicons!



Some Statistics

Currently!

Category	Lexical Concepts	Lexical entries	Synsets	Translations pairs	Glosses	Semantic relations
Total (Millions)	1.1 M	2.4 M	1.8 M	1.5 M	0.7 M	0.5 M
Sub Counts		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

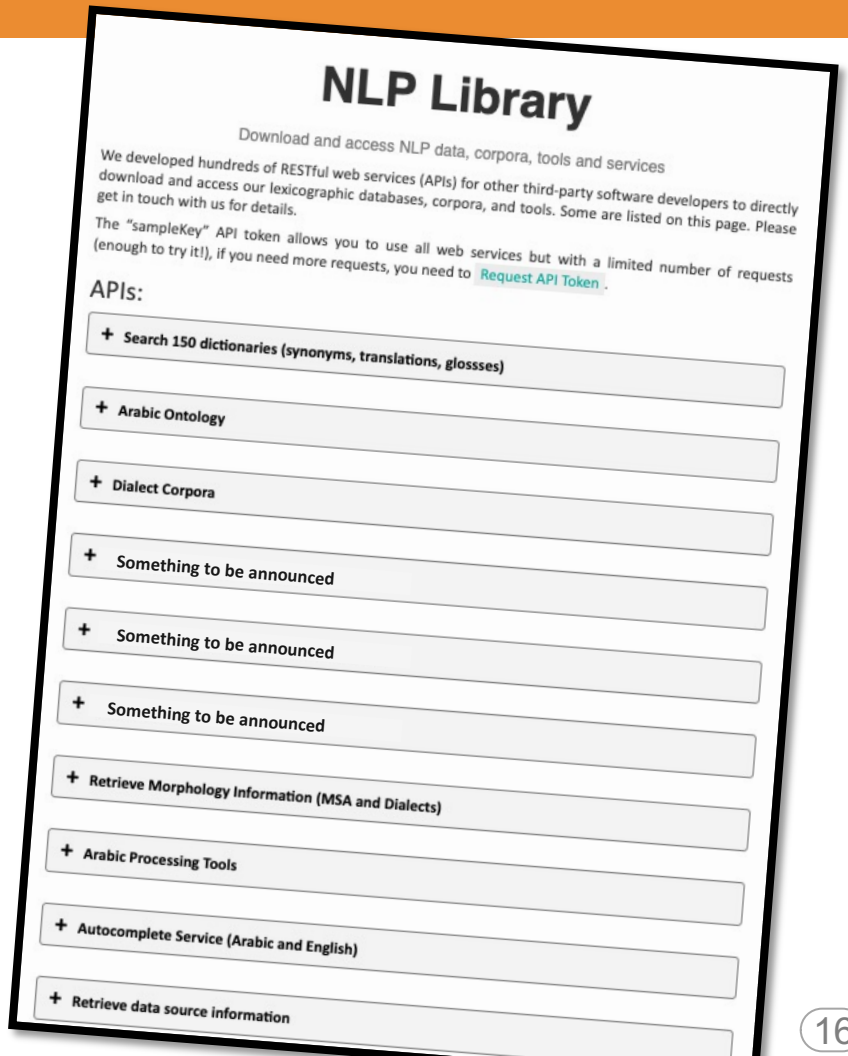
For more, see [3]

Accessible

To download and access NLP data, corpora, tools and services

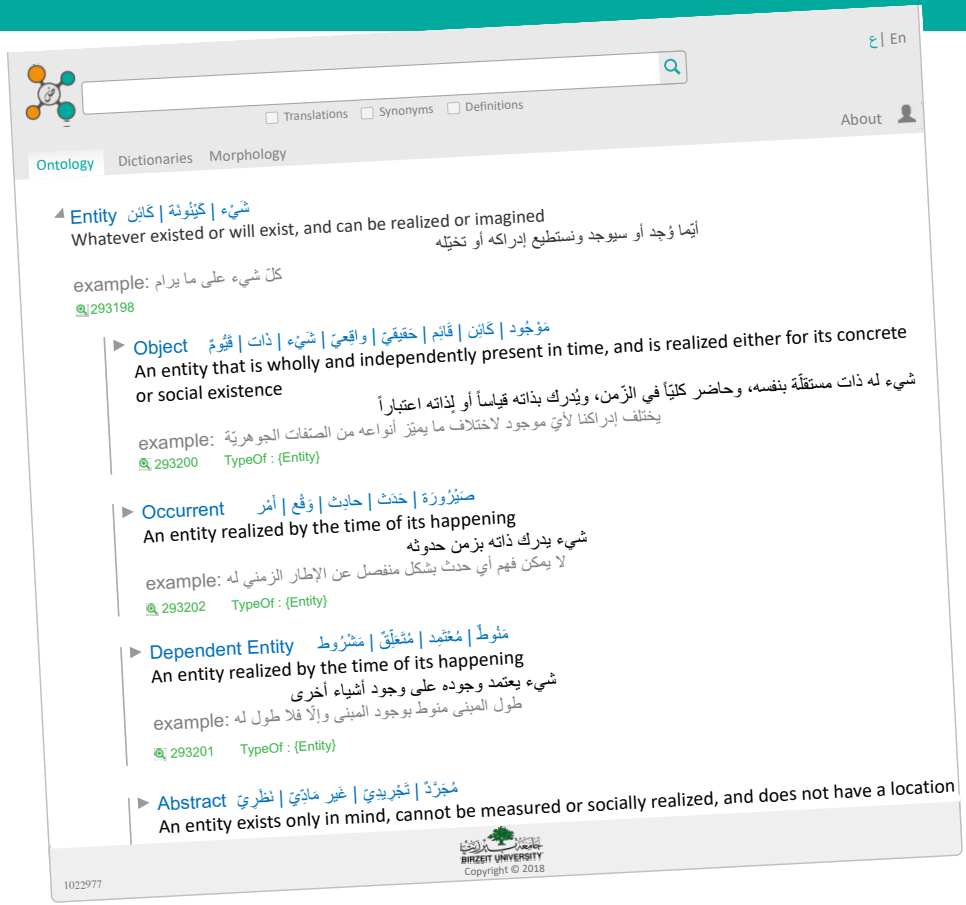
RESTful web services

Ask us for an API Key!



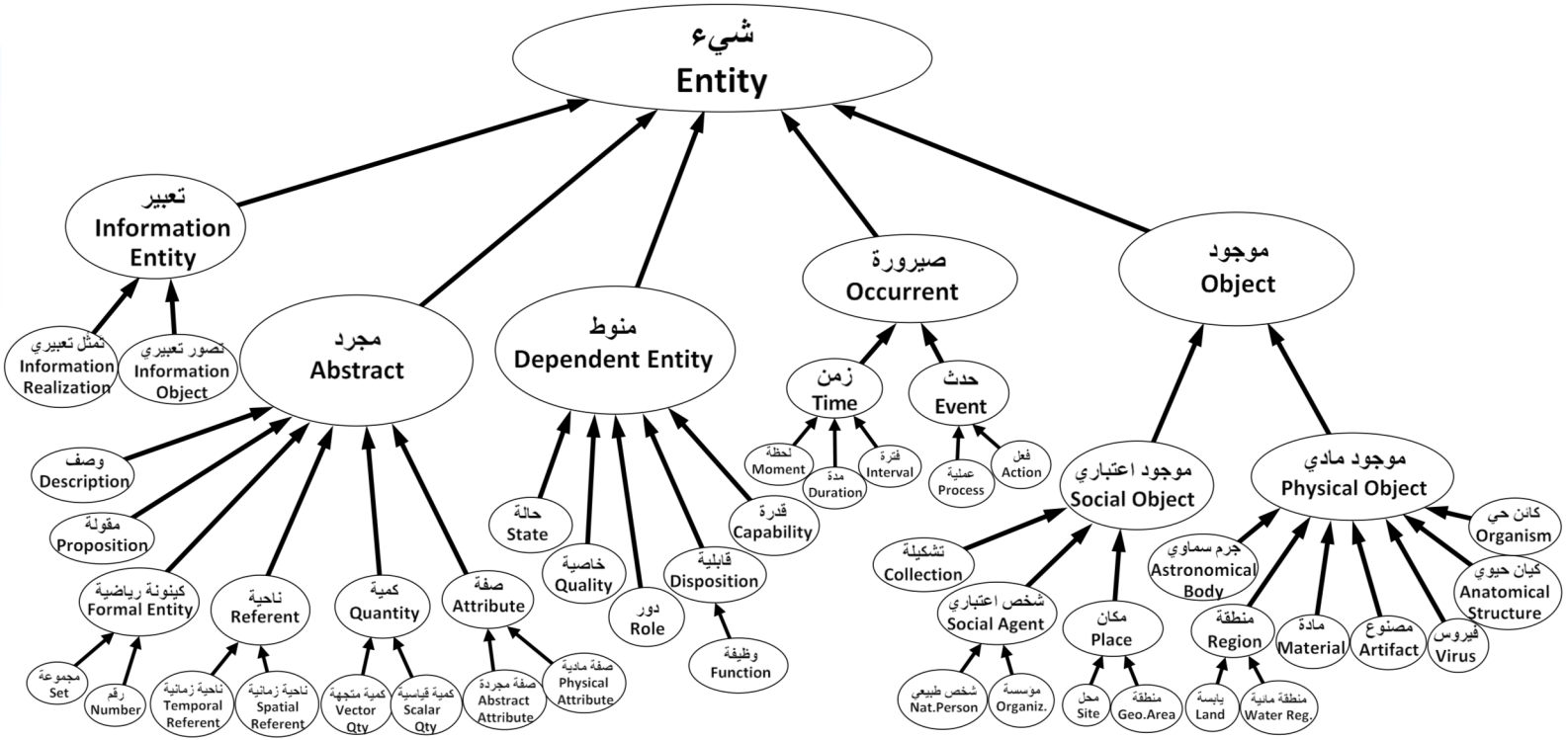
Arabic Ontology

- To enable machine understand semantics.
- Classification of the meanings of the Arabic terms, specified in D. Logic
- Benchmarked to scientific advances rather than to speakers' naïve beliefs as wordnets do.



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

Arabic Ontology

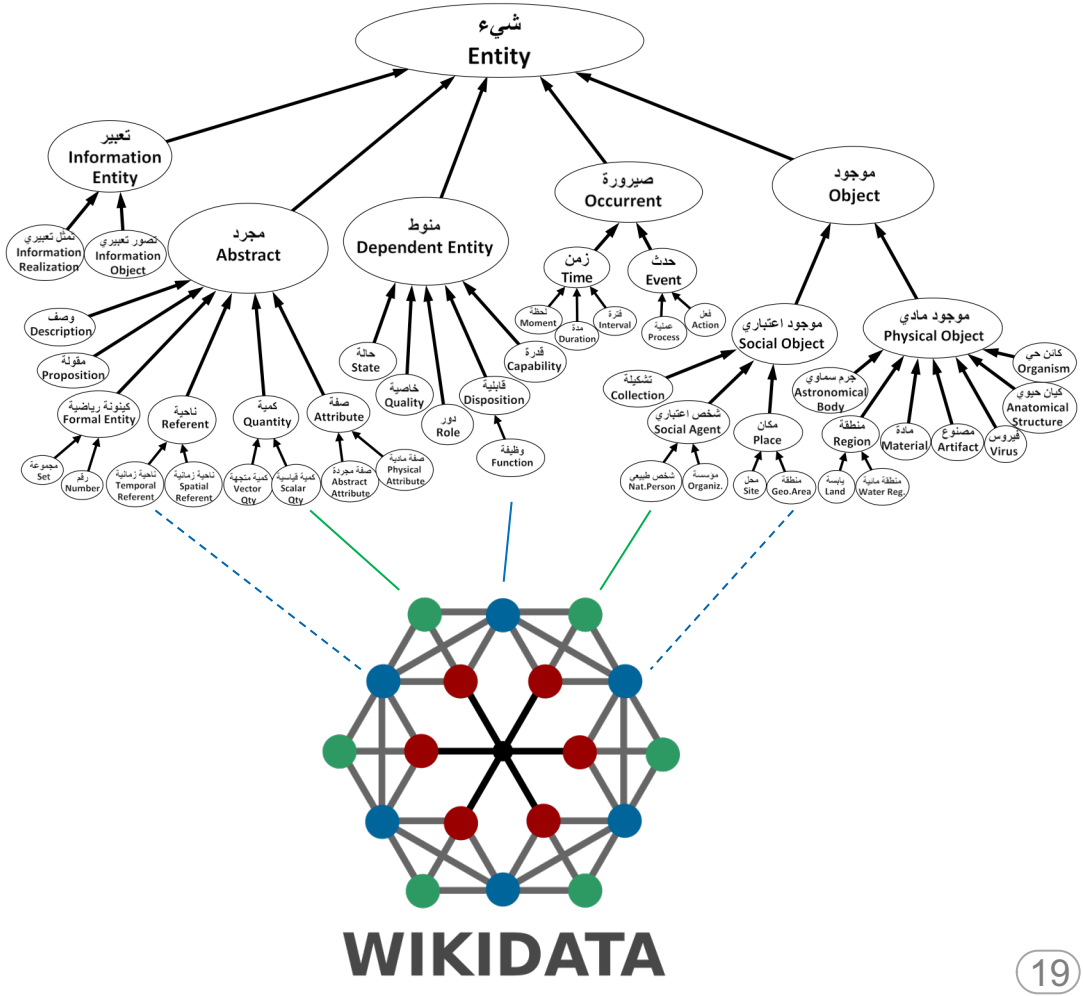


Based on:

Mustafa Jarrar: **The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content**. Applied Ontology Journal, IOS Press. (Forthcoming).

Linking the Ontology with Knowledge Graphs

- Every concept in the ontology is linked with a node in the Wikidata
- In this way, one can access the the world knowledge as an Arabic knowledge graph
- So, we can build applications like chatbots and question answering, …



Dialect Corpus

http://portal.sina.birzeit.edu/curras

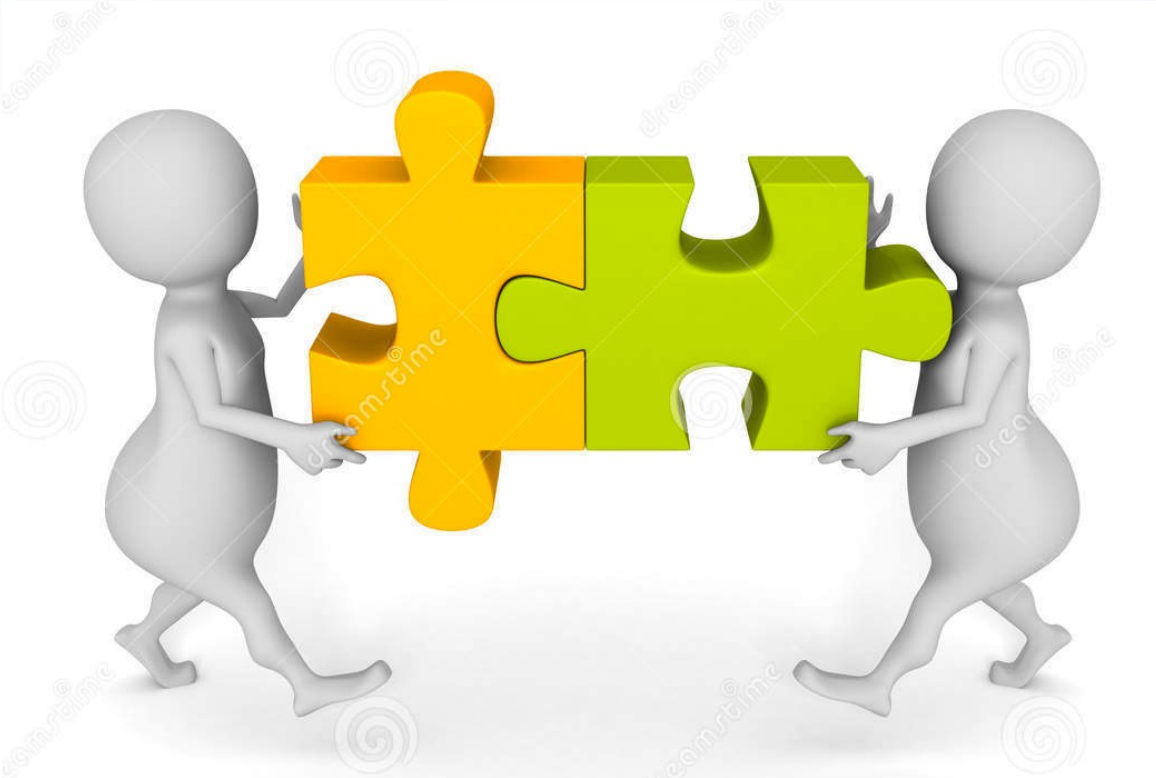
- To enable machines understand dialects.
- We collected a corpus written in Palestinian dialect (60k words).
- Described and annotate each word with 16 tags.

Based on:
Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, Nasser Zalmout: Curras: **An Annotated Corpus for the Palestinian Arabic Dialect.** Journal Language Resources and Evaluation. Volume(51), Issue(3). Springer. 2017

Dialect Corpus

http://portal.sina.birzeit.edu/curras

بنتنا زي كعك العيد بتنفق بسرعة. العريس : والله انتو عيلة ما بشتريكوا بقمعة سيجارة , اخر اشي بقولكوا اياه غروس مالبورو شو قلتوا ? الام : اسمع ولا , خاي ا	M	1	Verb	said;say;be_said; (for_example)_[CALIMA]	b/PROG_PART+A/IV 1S+qwl/IV+kw/IVSUF F_DO:2P	باقولكو	قال 1	قول	بقولكوا
زيون يا بقتل امي . منال : عنا بتطلع بالتكسي . . . ايه ازمة . بتسأل الشوفير شوف س ليش مأزمة ? بقولك ما هو اليوم الخميس هيك . بتطلع السبت يا زلمة شو القص	M	3	Verb	said;say;be_said; (for_example)_[CALIMA]	b/PROG_PART+y/IV 3MS+qwl/IV+k/IVSUF F_DO:2MS	بيقولك	قال 1	قول	بقولك
ير شوف س ليش مأزمة ? بقولك ما هو اليوم الخميس هيك . بتطلع السبت يا زلمة شو القصة ليش مأزمة ? بقولك يوم السبت عادة هيك , بتطلع الاثنين له له له شوه الأز	M	3	Verb	said;say;be_said; (for_example)_[CALIMA]	b/PROG_PART+y/IV 3MS+qwl/IV+k/IVSUF F_DO:2MS	بيقولك	قال 1	قول	بقولك
مة شو القصة ليش مأزمة ? بقولك يوم السبت عادة هيك , بتطلع الاثنين له له له شوه الأزمة شو في ? بقولك يوم الاثنين فش مجال دايما هيك . بتطلع الثلاثاء ول لي	M	3	Verb	said;say;be_said; (for_example)_[CALIMA]	b/PROG_PART+y/IV 3MS+qwl/IV+k/IVSUF F_DO:2MS	بيقولك	قال 1	قول	بقولك
الثلاثاء ول ليش كل هالأزمة . يوم الثلاثاء يختي معروف فس نفس . بتطلع الخميس ليش هالقدة الأزمة بقولك يوم الخميس دايما هيك , هذا يوم في الاسبوع طاير . بترو	M	3	Verb	said;say;be_said; (for_example)_[CALIMA]	b/PROG_PART+y/IV 3MS+qwl/IV+k/IVSUF F_DO:2MS	بيقولك	قال 1	قول	بقولك
اط@@صنّند @@Lاط@@ة منال. عنا بتطلع بالتكسي . . . ايه ازمة . بتسأل الشوفير شوف س ليش مأزمة ? بقولك ما هو اليوم الخميس هيك . بتطلع السبت يا زلمة شو القص	M	3	Verb	said;say;be_said; (for_example)_[CALIMA]	b/PROG_PART+y/IV 3MS+qwl/IV+k/IVSUF F_DO:2MS	بيقولك	قال 1	قول	بقولك
الثلاثاء ول ليش كل هالأزمة . يوم الثلاثاء يختي معروف فس نفس . بتطلع الخميس ليش هالقدة الأزمة بقولك يوم الخميس دايما هيك , هذا يوم في الاسبوع طاير . L@@	M	3	Verb	said;say;be_said; (for_example)_[CALIMA]	b/PROG_PART+y/IV 3MS+qwl/IV+k/IVSUF F_DO:2MS	بيقولك	قال 1	قول	بقولك



Connecting lexical resources

W3C Lemon RDF Model Standard

Lexicon Model for Ontologies: Community Report, 10 May 2016

Final Community Group Report 10 May 2016

Editors:

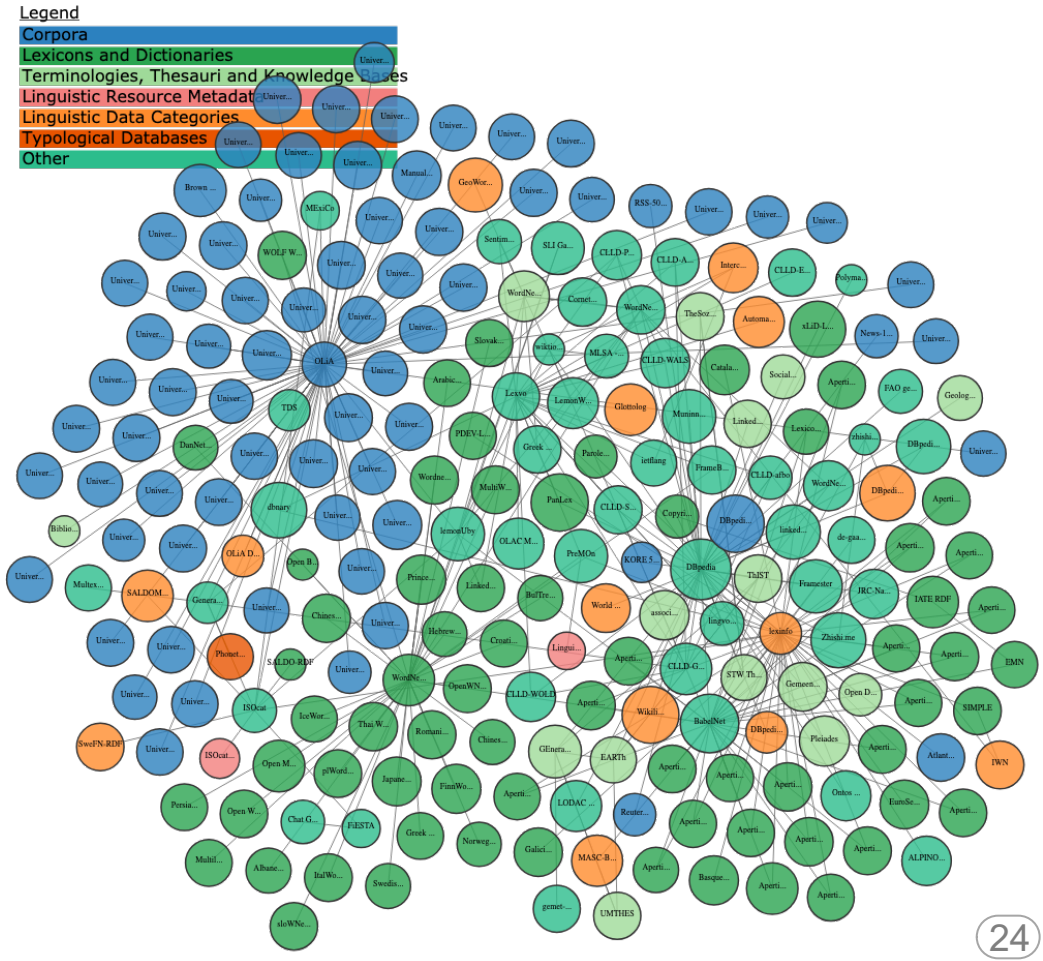
Philipp Cimiano (Cognitive Interaction Technology Excellence Center, Bielefeld University)
John P. McCrae (Insight Centre for Data Analytics, National University of Ireland, Galway)
Paul Buitelaar (Insight Centre for Data Analytics, National University of Ireland, Galway)

Abstract

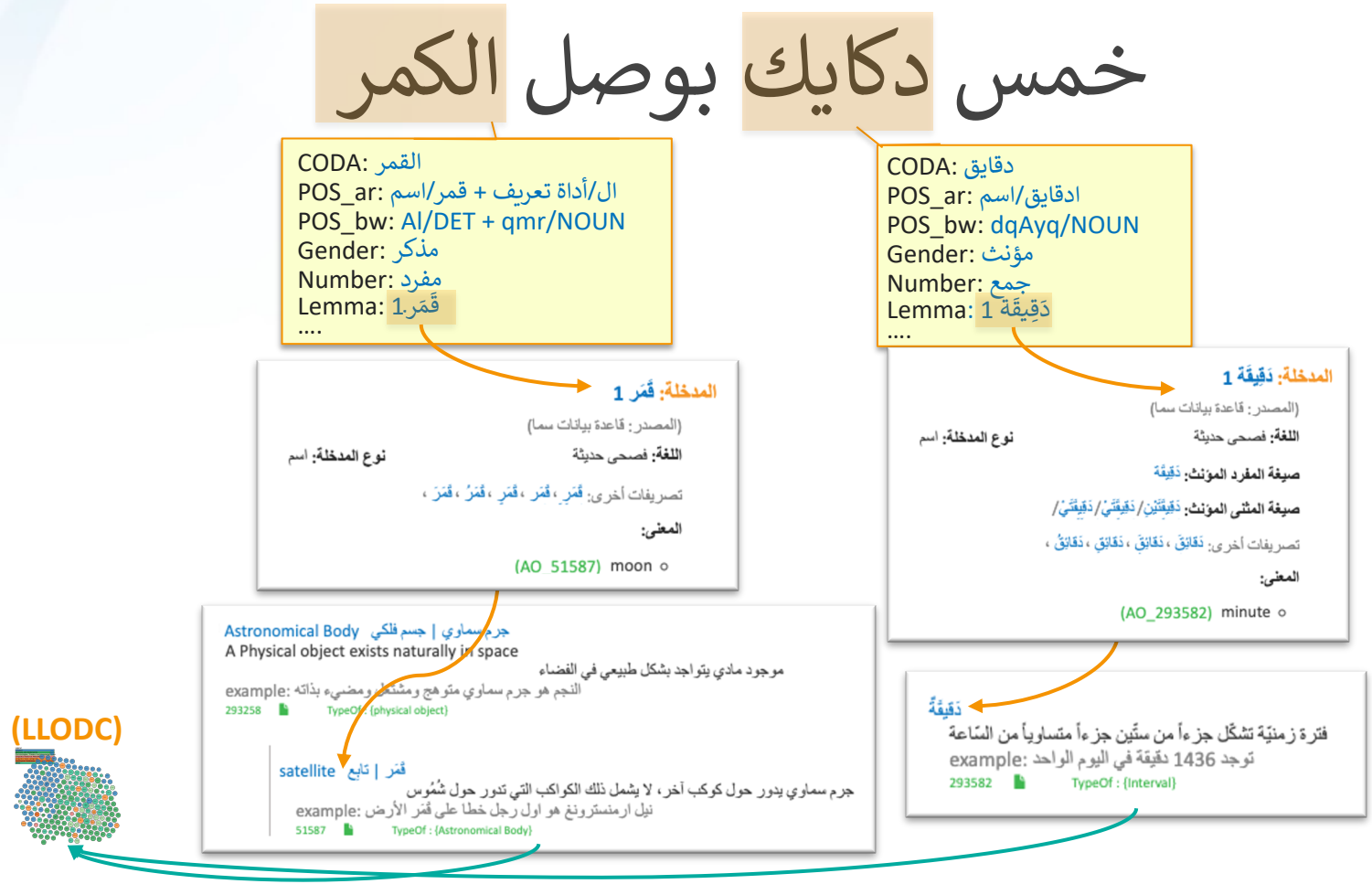
This document describes the lexicon model for ontologies (*lemon*) as a main outcome of the work of the Ontology Lexicon (Ontolex) community group.

The Linguistic Linked Open Data Cloud

- A collaborative effort to develop a Linked Open Data (sub-)cloud of linguistic resources.
- Represent (lexical entries, concepts, synsets, and other) using Lemon RDF model, then interlinked.



Linking All linguistics levels



Example Case Study

Word Sense Disambiguation

Moustafa Al-Hajj, Mustafa Jarrar: **ArabGlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD.** In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). PP 40--48, 2021

The problem

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 كعَيْنَيْ الصقر - ألا إنّما العينان للقلب رائدٌ ...
2. **جاسوس**، "كان عينًا لدولةٍ أجنبيَّة . بثَّ العيون : تجسَّس، راقب - فلانٌ عَيْن على فلان : ناظر عليه
3. **أجود كلّ شيء** وأحسنه ونفيسُه: عيونُ الفنِّ.
4. **حارس:** فلان عين على المكان.
5. **الحاضر من كل شيء** أصبح أثرًا بعد عين ...
6. **عَيْنُ الماء**:- ينبوعه، تُحلِّق الطّيورُ فوق عيون الماء
7. **عَيْن الشَّيء**:- نفسه، ذاته (تستعمل للتوكيد): جاء القوم أعينهم - كنّا في المكان عينه.
8. **عَيْن العقل**:- قدرة ذهنيَّة مَوْروثة على التخيُّل وتذكّر الأحداث.
9. .....

WSD has been a challenging task for many years but has gained recent attention due to the advances in contextualized word embedding models such as BERT.

Our Contributions

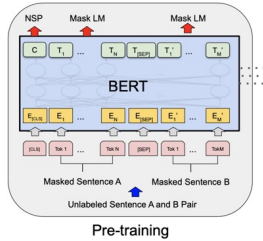
❖ Arabic context-gloss pairs Dataset (167k)

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

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

❖ Three Fine-tuned BERT Models

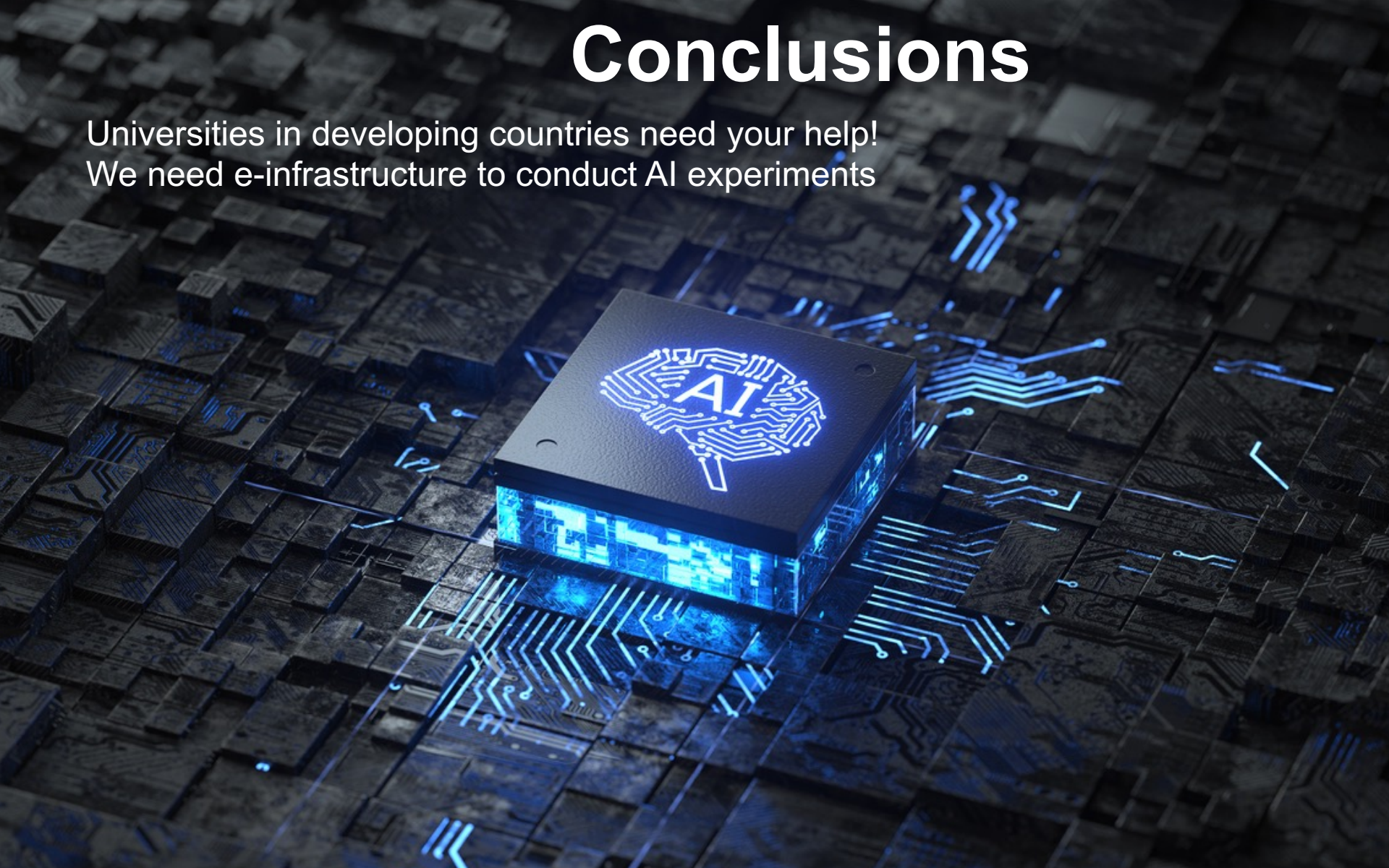
- WSD into **binary sequence-pair classification task**
- Accuracy 84%
- 4 types of signals to emphasize target words in context



Pre-training

Conclusions

Universities in developing countries need your help!
We need e-infrastructure to conduct AI experiments



References

1. Moustafa Al-Hajj, Mustafa Jarrar: **ArabGlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD.** In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). PP 40--48, 2021

2. Mustafa Jarrar. **The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content.** Applied Ontology Journal, IOS Press, 2021.

3. Mustafa Jarrar, Hamzeh Amayreh. **An Arabic-Multilingual Database with a Lexicographic Search Engine.** Proceedings of the 24th International Conference on Applications of Natural Language to In- formation Systems (NLDB), 2019.
Mustafa Jarrar: **Search Engine for Arabic Lexicons.** Proceedings of the 5th Conference on Translation and the Problematics of Cross-cultural Understanding. The Forum for Arab and International Relations. Doha, Qatar. 2018

4. Hamzeh Amayreh, Mohammad Dwaikat, and Mustafa Jarrar. **Lexicon Digitization -A Framework for Structuring, Normalizing and Cleaning Lexical Entries**. Technical Report, Birzeit University. 2018.

5. Mustafa Jarrar, Fadi Zaraket, Rami Asia, Hamzeh Amayreh: **Diacritic-Based Matching of Arabic Words**. ACM Asian and Low-Resource Language Information Processing. Volume 18, No 2, Pages(10:1--10:21), ACM, December 2018.

6. Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. **Curras: An Annotated Corpus for the Palestinian Arabic Dialect**. Journal Language Resources and Evaluation, 51(3):745–775, 2017.

7. Mustafa Jarrar, Nizar Habash, Diyam Akra, Nasser Zalmout: **Building a Corpus for Palestinian Arabic: a Preliminary Study.** In proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing. Association for Computational Linguistics (ACL), Pages (18-27). October 25, 2014, Doha, Qatar. ISBN: 978-1-937284-96-1

8. Mustafa Jarrar. **Building a Formal Arabic Ontology (Invited Paper).** In Proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks. ALECSO, Arab League, 2011

Thank You

Mustafa Jarrar
mjarrar@birzeit.edu