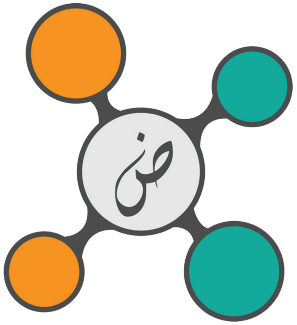
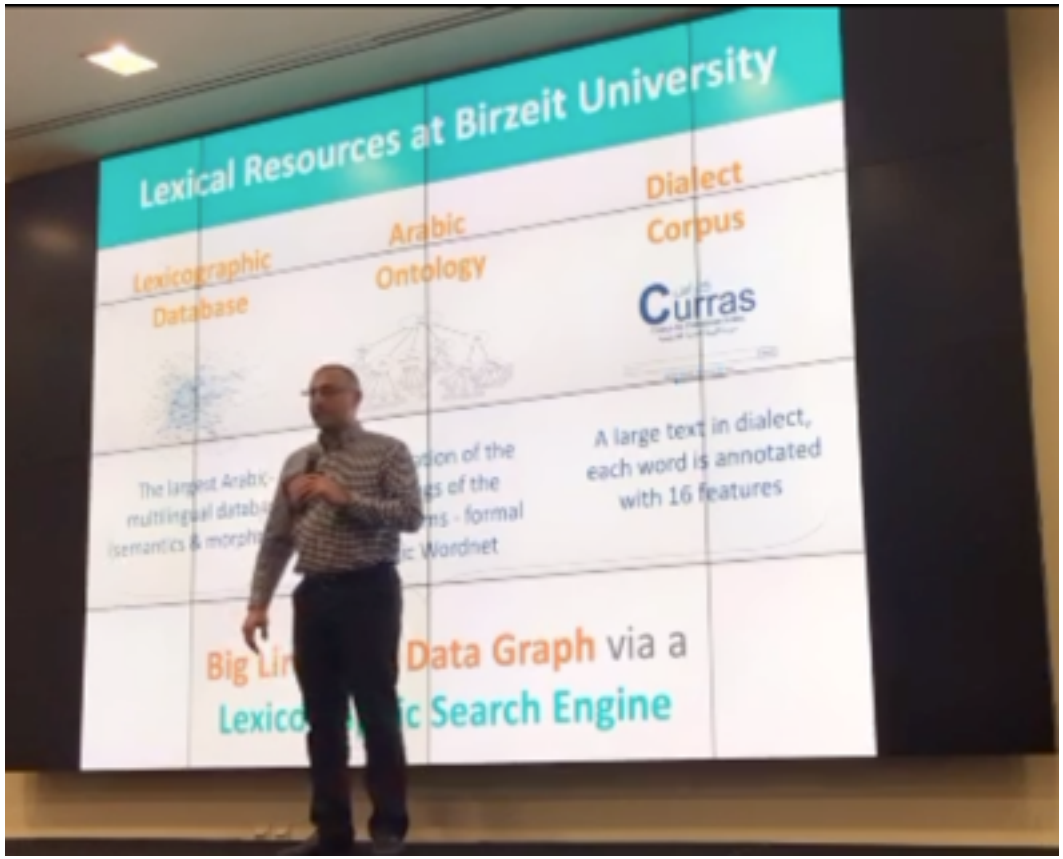
The 1st Arabic Speech Workshop.
Doha, April 2019



Towards building a
Big Linguistic Data Graph

Mustafa Jarrar
Birzeit University
Palestine

Watch the lecture here



https://twitter.com/ArabicSpeech/status/1122843057756475392?s=20

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!

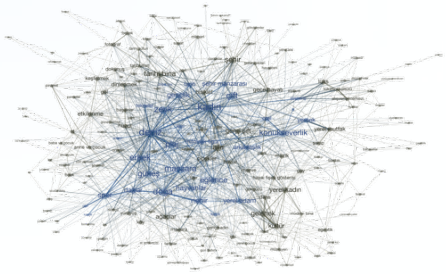
Our Solution

Digitize, Collect, Build, then clean and link

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

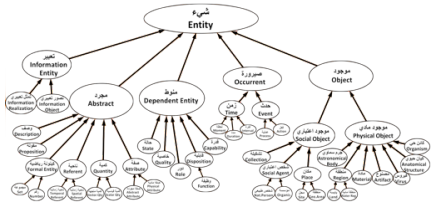
Lexical Resources at Birzeit University

Lexicographic Database



The largest Arabic-multilingual database (semantics & morphology)

Arabic Ontology



Classification of the meanings of the Arabic Terms - formal Arabic Wordnet

Dialect Corpus

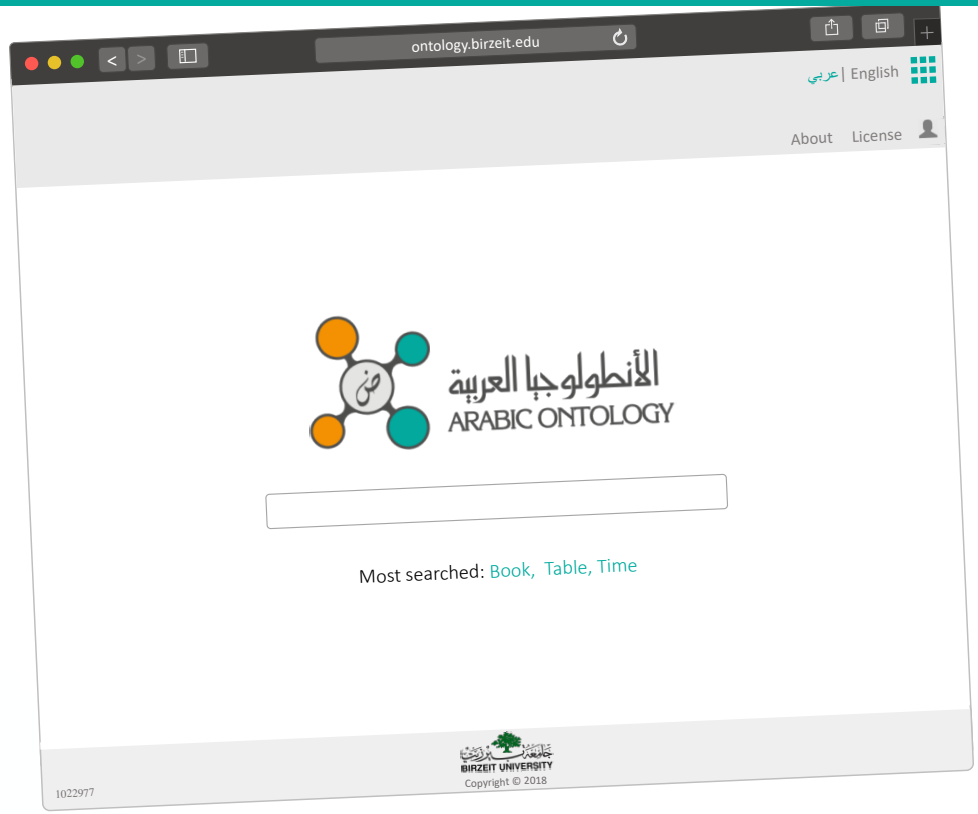
A large text in dialect, each word is annotated with 16 features

Big Linguistic Data Graph via a **Lexicographic Search Engine**

Lexicographic Database

Lexicographic Search Engine

- **Free access to people**: students, translators, researchers, Arabic learners …
- **API accessible** for NLP applications.



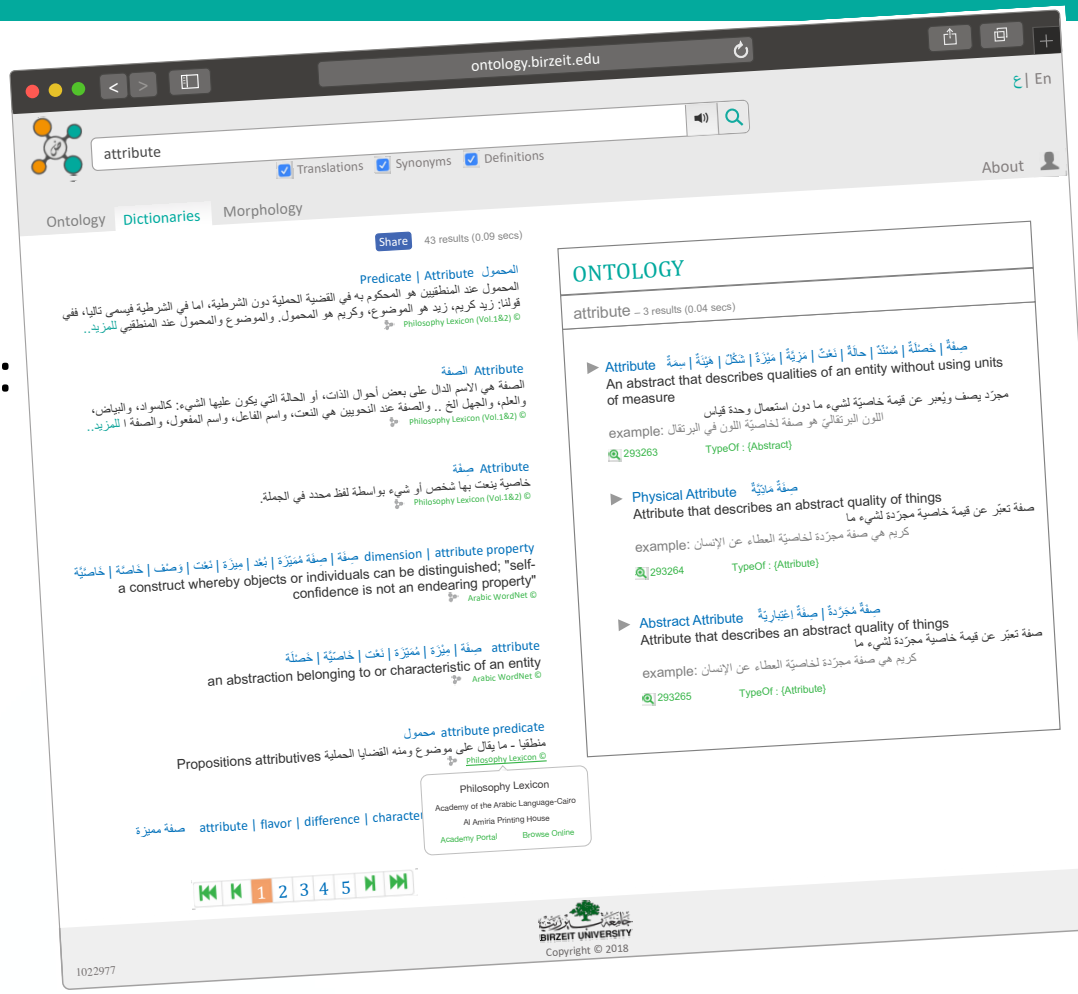
https://ontology.birzeit.edu

Based on:

Mustafa Jarrar, Hamzeh Amayreh: **An Arabic Multilingual Database with a Lexicographic Search Engine**. The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019), Springer. Manchester, UK. 2019

Lexicographic Search Engine

- **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

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

DEMO

https://ontology.birzeit.edu

Lexicographic Search Engine

Conformance with W3C Standards:

- ✓ **W3C's Best Practices for Publishing Linked Data**
  including the Cool URIs, simplicity, stability, and linking

- ✓ **W3C's RDF Lemon Model**

التسوية levelling | grading

تحريك التربة أثناء إعداد الأرض للري للوصول إلى سطح مستو أو سطح ذي انحدار منتظم.

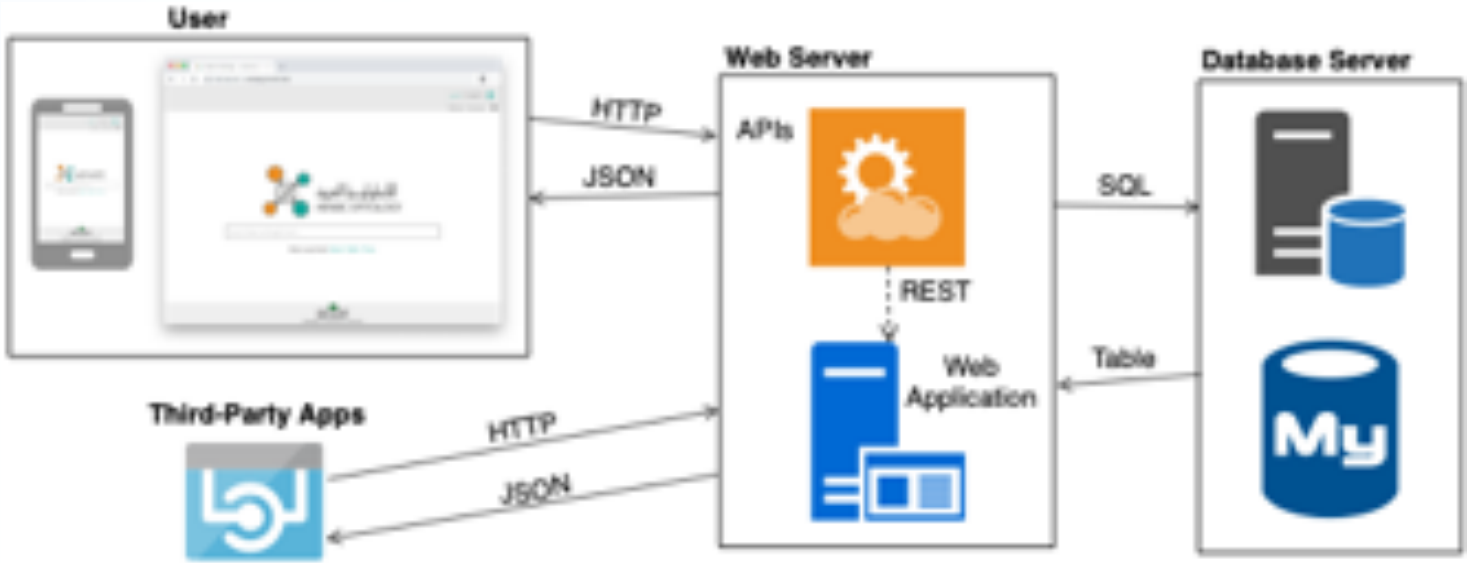
Hydrology Lexicon ©

```
...
@prefix aot: <http://ontology.birzeit.edu/term/>.
@prefix aoc: <http://ontology.birzeit.edu/lexicalconcept/>.
@prefix aor: <http://ontology.birzeit.edu/lexicon/>.

<aoc:1623> a ontolex:LexicalConcept;
ontolex:isEvokedBy <aot:Lex-grading>;
ontolex:isEvokedBy <aot:Lex-levelling>;
ontolex:isEvokedBy <aot:Lex-تسوية>;
skos:definition "تحريك التربة أثناء إعداد الأرض للري للوصول إلى سطح مستو أو سطح..."@ar;
skos:inScheme <aor:Hydrology_Lexicon_1>.

<aot:lex-grading> a ontolex:LexicalEntry, ontolex:Word;
ontolex:canonicalForm [ontolex:writtenRep "grading"@en];
skos:inScheme <aor:Hydrology_Lexicon_1>.
<aot:lex-levelling> a ontolex:LexicalEntry, ontolex:Word;
ontolex:canonicalForm [ontolex:writtenRep "levelling"@en];
skos:inScheme <aor:Hydrology_Lexicon_1>.
<aot:lex-تسوية> a ontolex:LexicalEntry, ontolex:Word;
ontolex:canonicalForm [ontolex:writtenRep "تسوية"@ar];
skos:inScheme <aor:Hydrology_Lexicon_1>.
```

Based On:

Mustafa Jarrar, Hamzeh Amayreh, John McCarae: **Progress on Representing Arabic Lexicons in Lemon**. The 2nd Conference on Language, Data and Knowledge (LDK 2019), Germany. 2019.

Lexicographic Search Engine



Search Engine Architecture

Lexicographic Search Engine

Some Statistics

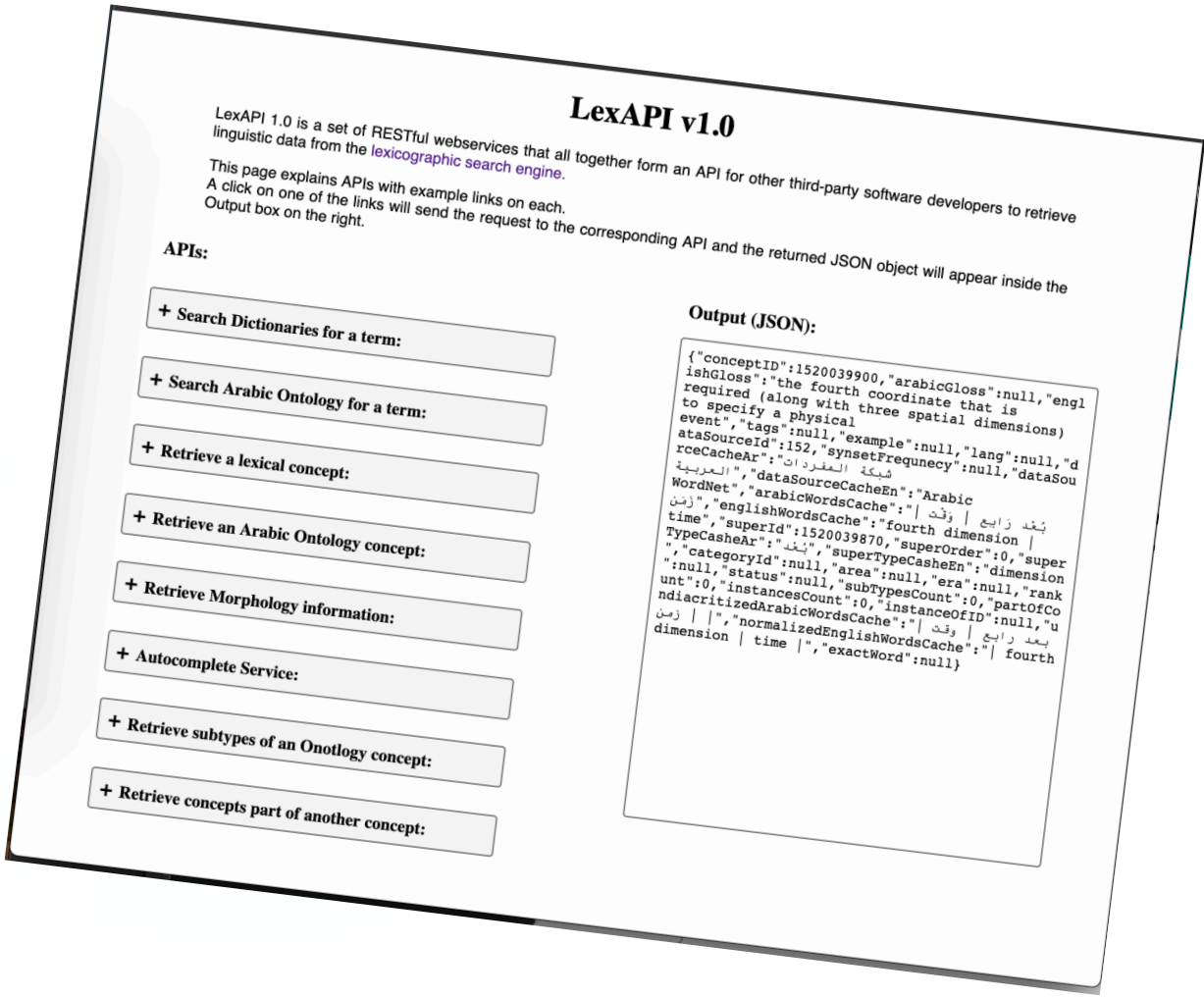
Category	Lexical Concepts	Lexical entries	Synsets	Translations pairs	Glosses	Semantic relations
Total (Millions)	1.1 M	2.4 M	1.8 M	1.5 M	0.7 M	0.5 M
Sub Counts		1,100K Arabic 1,100K English 200K French 3K Others 1,300K Single-word 1,000K Multi-word	800K Arabic 800K English 200K French 50K Others	1,000K 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

Lexicographic Search Engine

API Access

RESTful web services

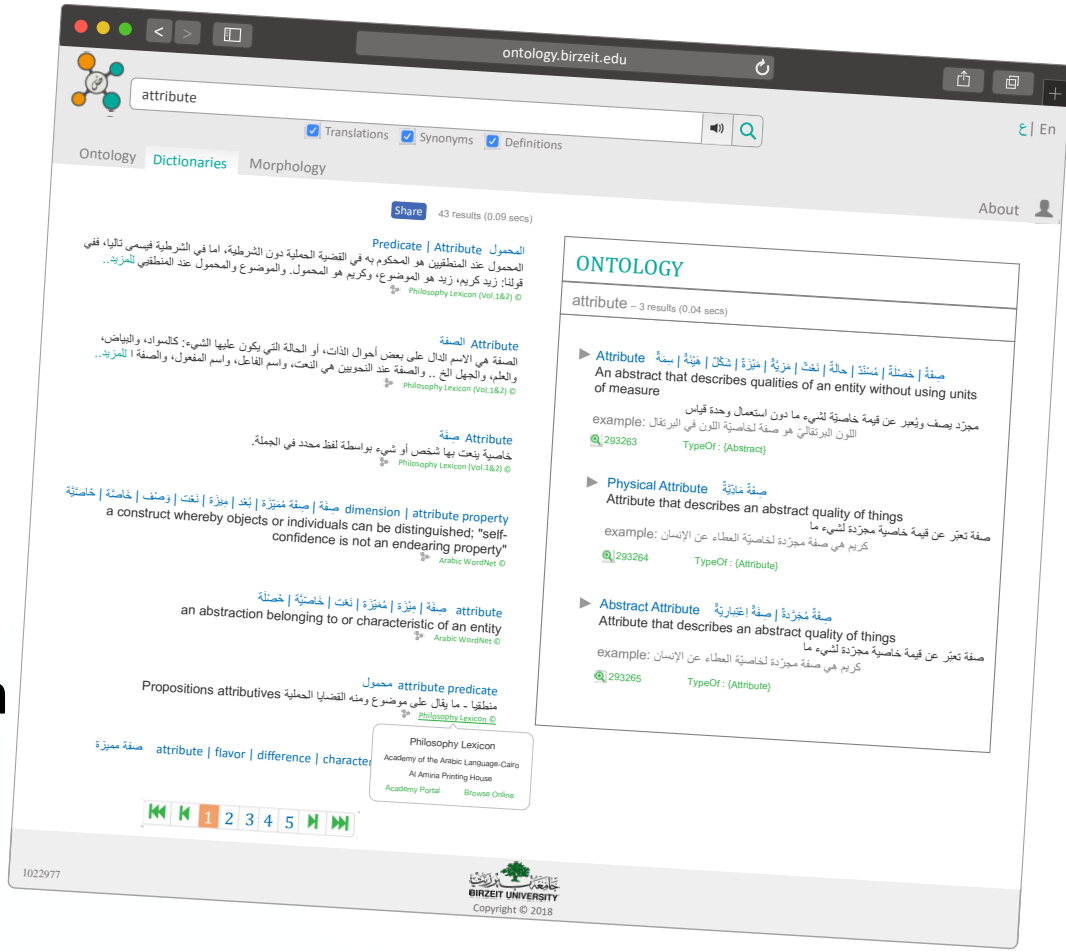
Ask us for an API Key!



Lexicographic Search Engine

Done over 9 years:

- Obtain permission from each dictionary owner.
- Manually type, then structure and store into one database.
- Didn’t receive any fund from any party!
- Non-profit research

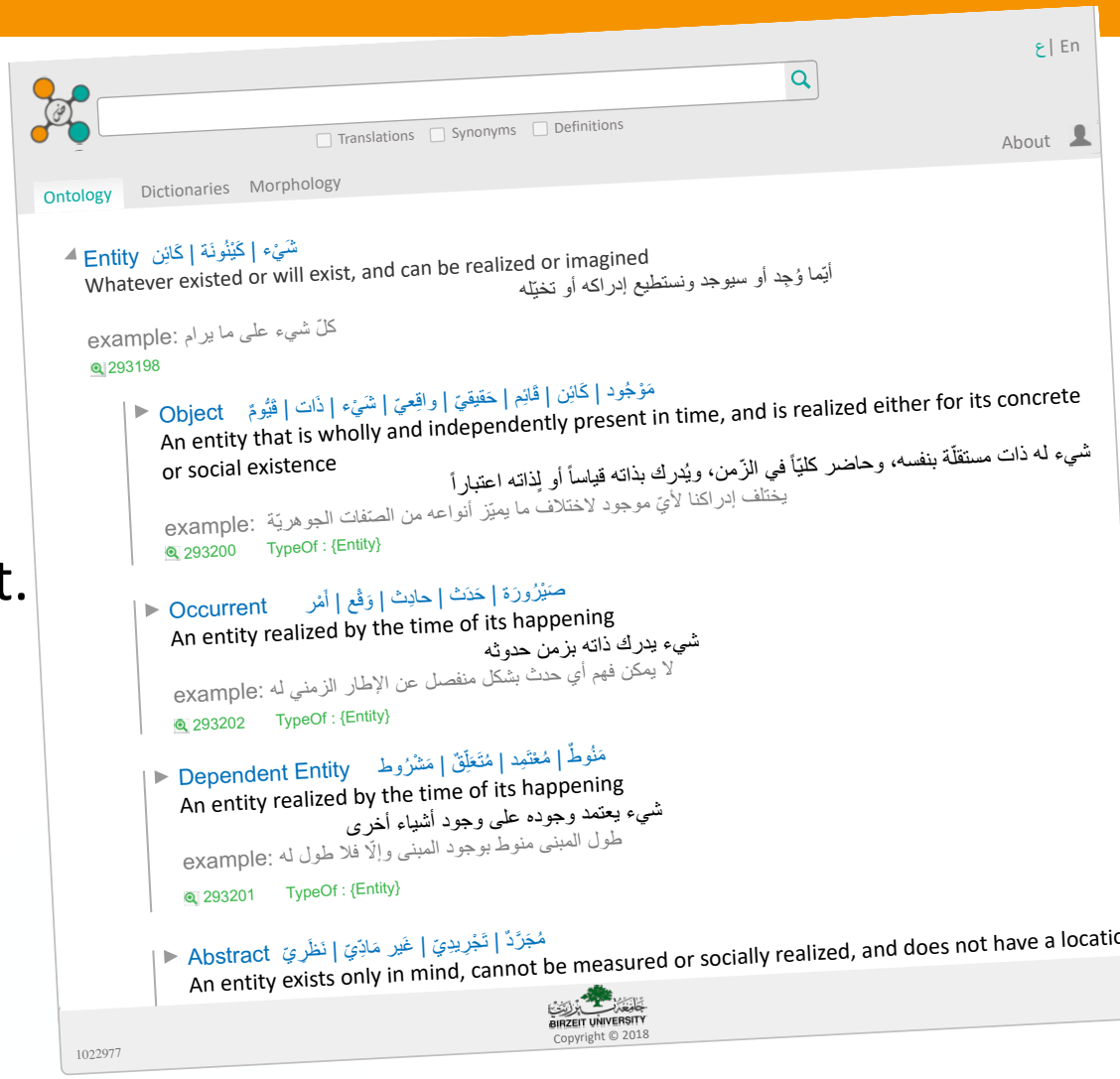


https://ontology.birzeit.edu

Arabic Ontology

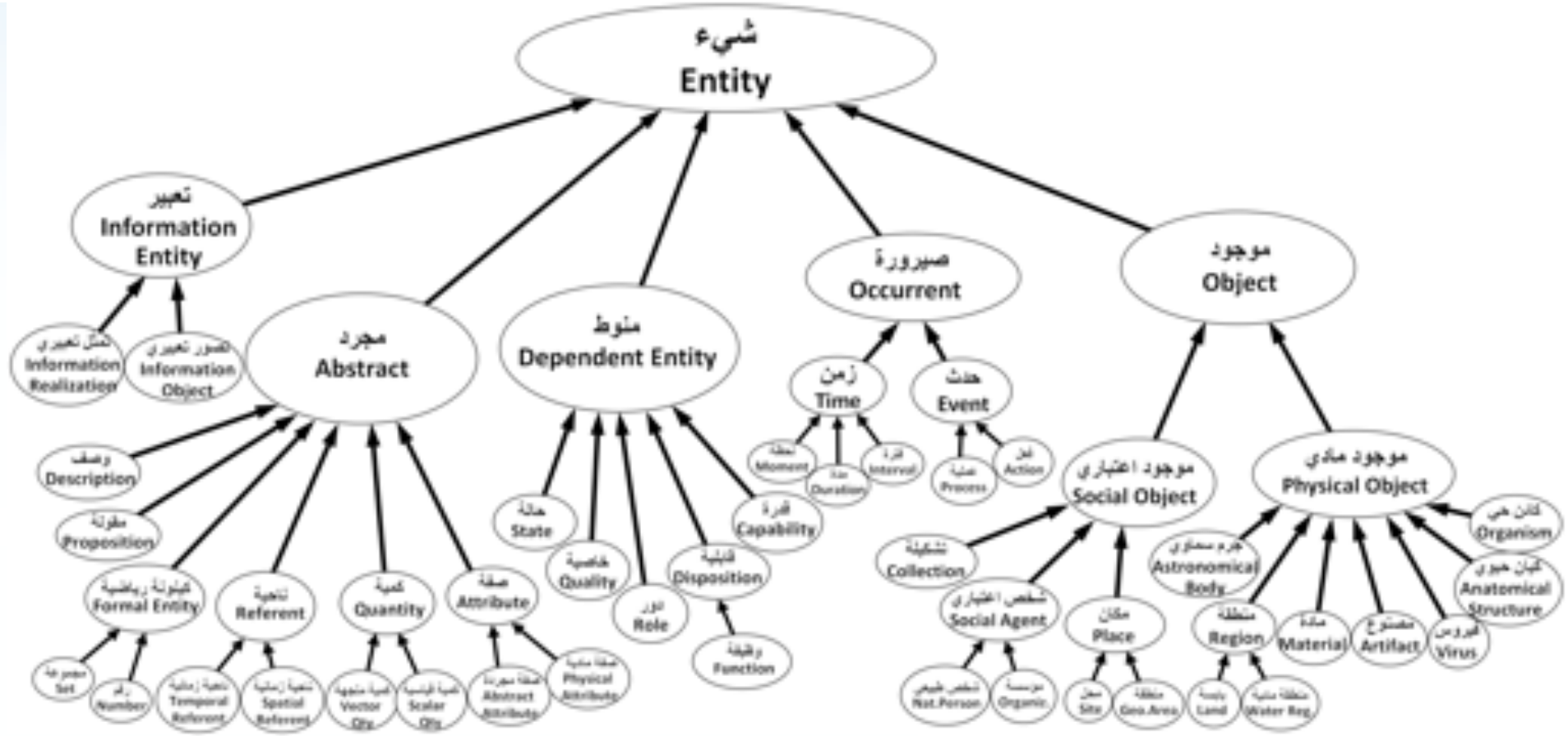
Arabic Ontology

- Classification of the meanings of the Arabic terms, specified in D. Logic
- Can be used as a formal Arabic Wordnet -with ontologically-clean content.
- Linked with WordNet
- 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).

Dialect Corpus

Dialect Corpus

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

- Collect a corpus written in Palestinian dialect.
- Describe and annotate each word with 16 tags.
- Stats:
  60k tokens,
  16K unique tokens

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

Dialect Corpus

Size

Middle size!

Resource	Tokens (%)	Documents
Facebook	4852 (8.6)	35 threads
Twitter	4694 (8.3)	38 threads
Blogs	11,245 (19.8)	37 threads
Forums	1027 (1.8)	33 threads
Palestinian stories	3149 (5.6)	6 stories
Palestinian terms	1468 (2.6)	1 doc
TV Shows: وطن ع وتر *Watan Aa Watar*	30,265 (53.4)	41 episodes
Total	56,700	190

Accuracy of annotators

Accurate

Category	A1 accuracy	A2 accuracy
Complex-prefix	97.4	97.6
Stem	86.3	93.1
Complex-suffix	92.5	96.0
Surface (CODA)	90.8	97.8
Lemma	93.4	93.1
$\text{Lemma}_{\text{MSA}}$	87.5	87.2
Gloss	74.6	80.9
POS	94.1	95.7
Person	97.8	98.8
Aspect	99.3	99.1
Gender	86.2	96.9
Number	94.8	96.8

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:

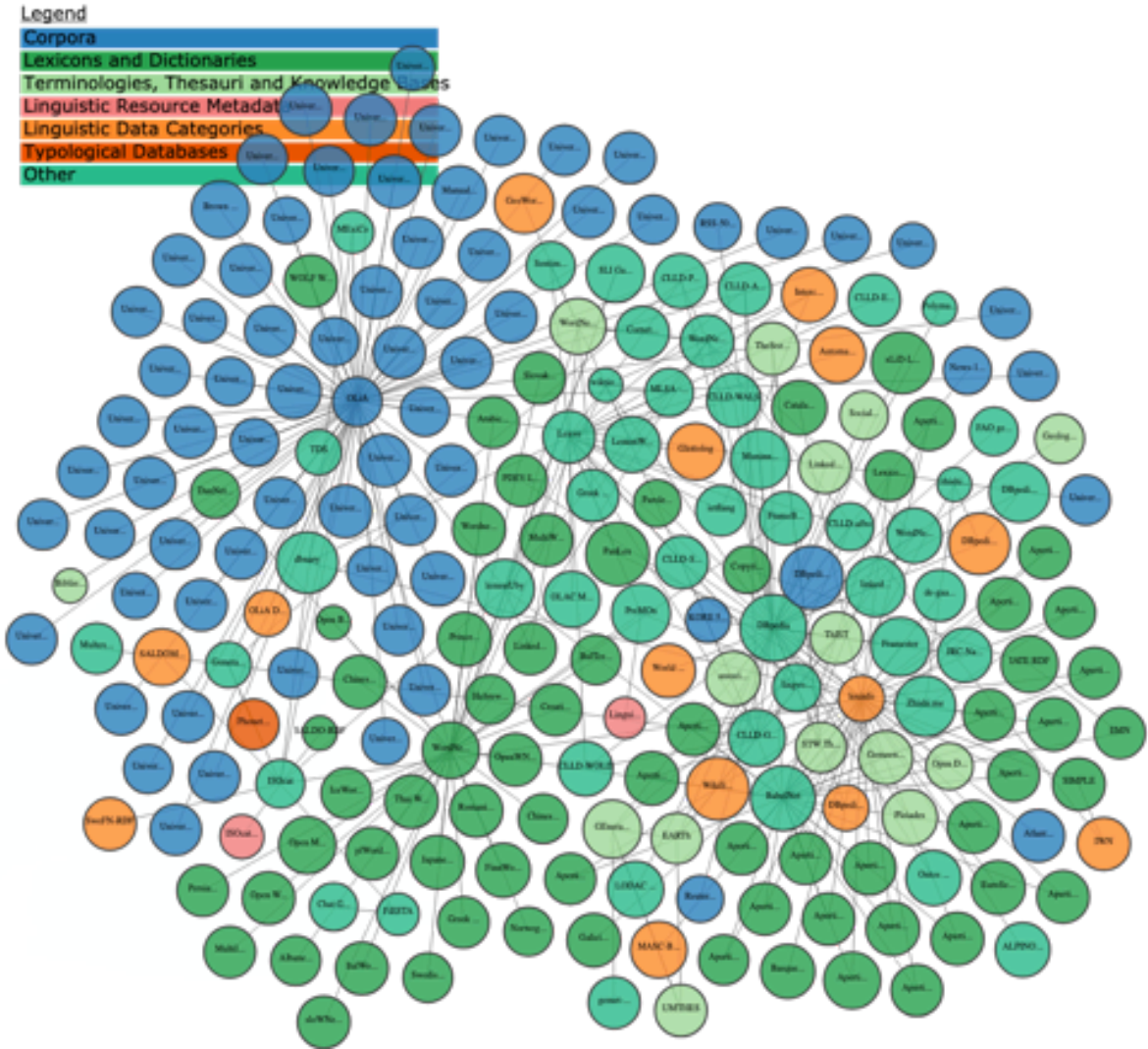
Philipp Cimiano (Cognitive Interaction Technology Excellence Center, Bielefeld University)
John P. McCrae (Insight Centre for Data Analytics, National University of Ireland, Galway)
Paul Buitelaar (Insight Centre for Data Analytics, National University of Ireland, Galway)

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.



Work in Progress

- Lemon-izing and Interlinking Arabic resources with the Linguistic Linked Open Data Cloud
- Building an Arabic Knowledge Graph
- Collect and interlink dialect corpora

References

1. Mustafa Jarrar. The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. Applied Ontology Journal, 2019 [Forthcoming].
2. Mustafa Jarrar and Hamzeh Amayreh. An Arabic-Multilingual Database with a Lexicographic Search Engine. Proceedings of the 24th International Conference on Applications of Natural Language to Information Systems (NLDB), 2019.
   Mustafa Jarrar, Hamzeh Amayreh: An Arabic Multilingual Database with a Lexicographic Search Engine. The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019), Springer. Manchester, UK. 2019
3. 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, 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