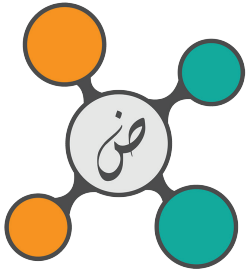


نحو بناء قواعد البيانات معجمية للفصحى والعامية

مصطفى جرار
جامعة بيرزيت - فلسطين

الذكاء الاصطناعي

فرصة للدول الصغيرة للتنافس والبروز

يمثل قلب الثورة الصناعية الرابعة

فرصة الاستثمار باللغة العربية صناعة وبحث علمي مستدام


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

سوق ضخمة جدا

٣٠٠ مليون
مستخدم في
الوطن العربي
(لغة أولى)

١,٥مليار
مستخدم للغة
العربية
(كلغة ثانية/دينية)

جهات وأسواق
أخرى
لدواعي اقتصادية
وأمنية

أثر عدم دعم التطبيقات الحديثة للعربية

على التعليم والثقافة والاقتصاد والسياسة وذوي الاجتياحات الخاصة....

لماذا نحتاج إلى مصادر لغوية

- ازدادت أهمية المصادر اللغوية (معاجم، مكانز، مسارد، أنطولوجيات..) لبناء التطبيقات الحاسوبية، مثل:
  - الترجمة الآلية
  - استرجاع البيانات بأكثر من لغة
  - فهم وتحليل النصوص المكتوبة والمنطوقة
  - التحدث مع الآلة

- شح في المصادر اللغوية العربية المحوسبة

  تخيلوا أنه لا يوجد قائمة بجميع المدخلات العربية!!!

- لا يوجد مصادر لغوية حديثة

  مثلاً، لا يوجد مكانز مترادفات تشمل أغلب الكلمات، أو مكانز مصممة بشكل جيد

➢ لهذا، يوجد ضعف في دعم التطبيقات للغة العربية

مساهمتنا في جامعة بيرزيت

جمع ورقمنة وتنقيح
ودمج وتوحيد المصادر اللغوية

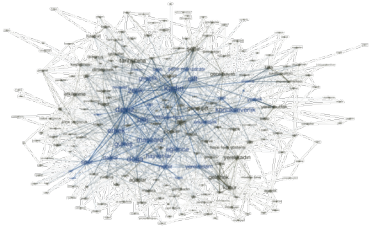
- وإتاحتها للجمهور العربي والباحثين ومتعلمي اللغة العربية
- وإتاحتها لمطوري التطبيقات عبر واجهات برمجية (APIs)

المصادر اللغوية في جامعة بيرزيت

مدونة للعامية

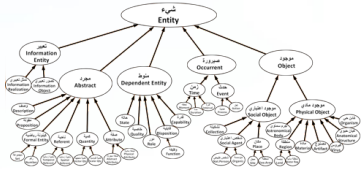
مجموعة ضخمة من النصوص العامية، تم تصريف ووسم كل كلمة فيها بحوالي 16 سمة.

قاعدة بيانات معجمية



أضخم قاعدة بيانات معجمية في تاريخ العربية (تصريفات، اشتقاقات، دلالة)

الأنطولوجيا العربية



تصنيف مفاهيم (معاني) الكلمات العربية
Formal Arabic Wordnet
عمل فلسفي حاسوبي لغوي

Linguistic Big Data Graph
شبكة بيانات لغوية ضخمة
تجمع المستويات التصريفية والاشتقاقية والدلالية للفصحى والعامية

https://ontology.birzeit.edu

قاعدة بيانات معجمية

Lexicographic Database

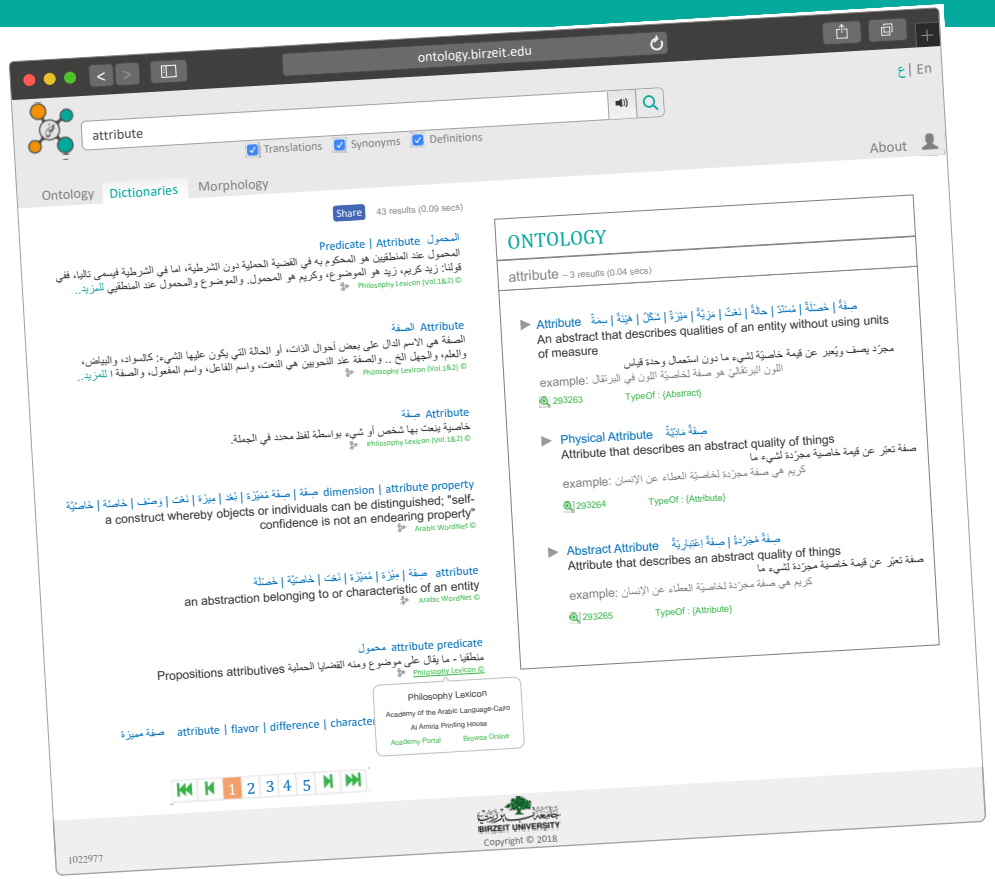
محرك بحث معجمي

- **البحث في 150 معجماً واسترجاع** تعريفات، ومترادفات، وفروق لغوية، وترجمات متخصصة.
- **دقيق!** مقارنة بالمترجمات الآلية.
- **الأول عالمياً،** فمثلاً، لا يوجد محرك بحث للإنجليزية، يحتوي هذا الكم من المعاجم في قاعدة بيانات واحدة.

https://ontology.birzeit.edu

قاعدة بيانات معجمية

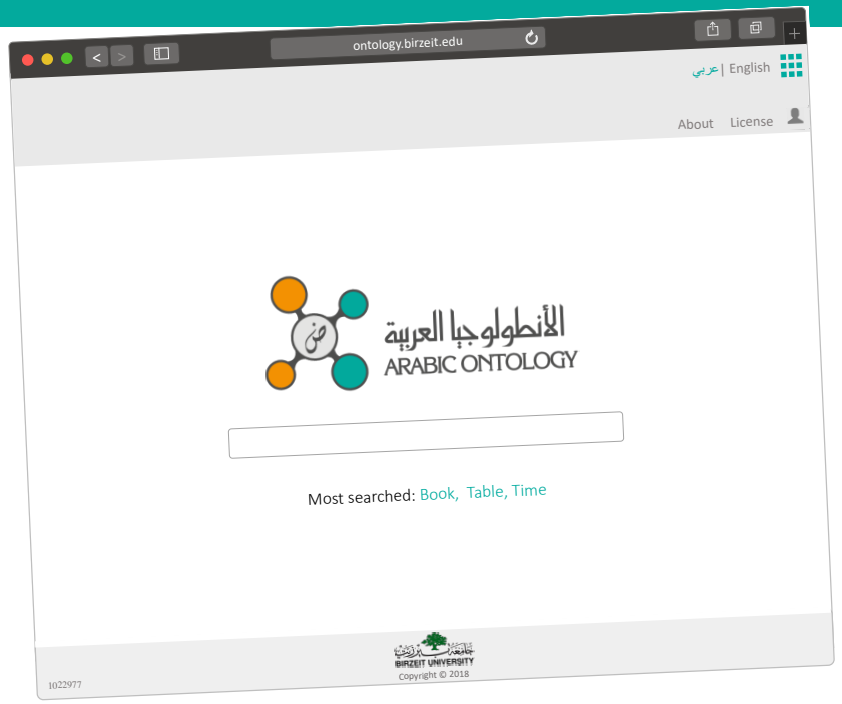
- **عربية ومتعددة اللغات**
- **تحتوي على أنواع كثيرة من المصادر اللغوية:** معاجم لغوية، ومعاجم ثنائية/ثلاثية اللغة، وقواعد بيانات تصريفية، ومسارد، ومكانز، وفروق لغوية، والأنطولوجيا العربية، وغيرها.
- **تغطي معظم المجالات:** علوم، وهندسة، وصحة وطب، وفلسفة، وإنسانيات، وفنون، ...



https://ontology.birzeit.edu

محرك بحث معجمي

- **متاحة للعامة:** طلبة، مترجمين، باحثين، متعلمي اللغة ...
- **متاح لمطوري التطبيقات:** عبر واجهات برمجية (APIs)



https://ontology.birzeit.edu

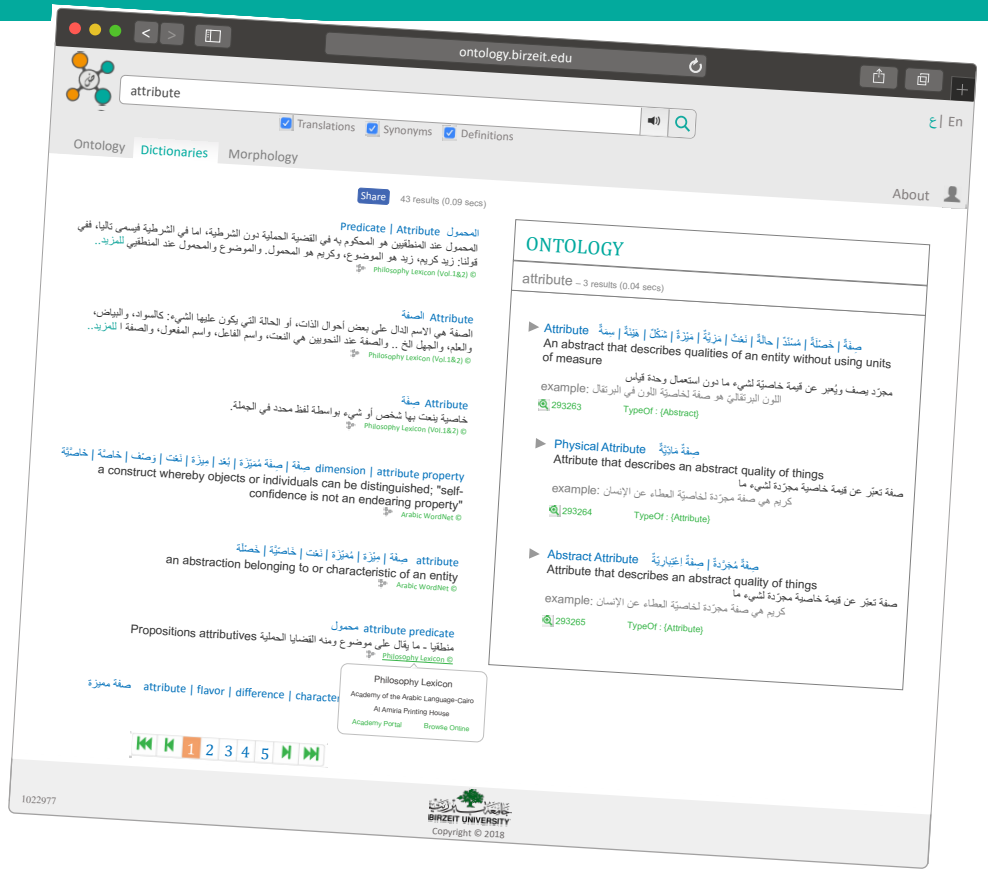
Based on:

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

محرك بحث معجمي

تم إنجازه عبر عشرة سنوات

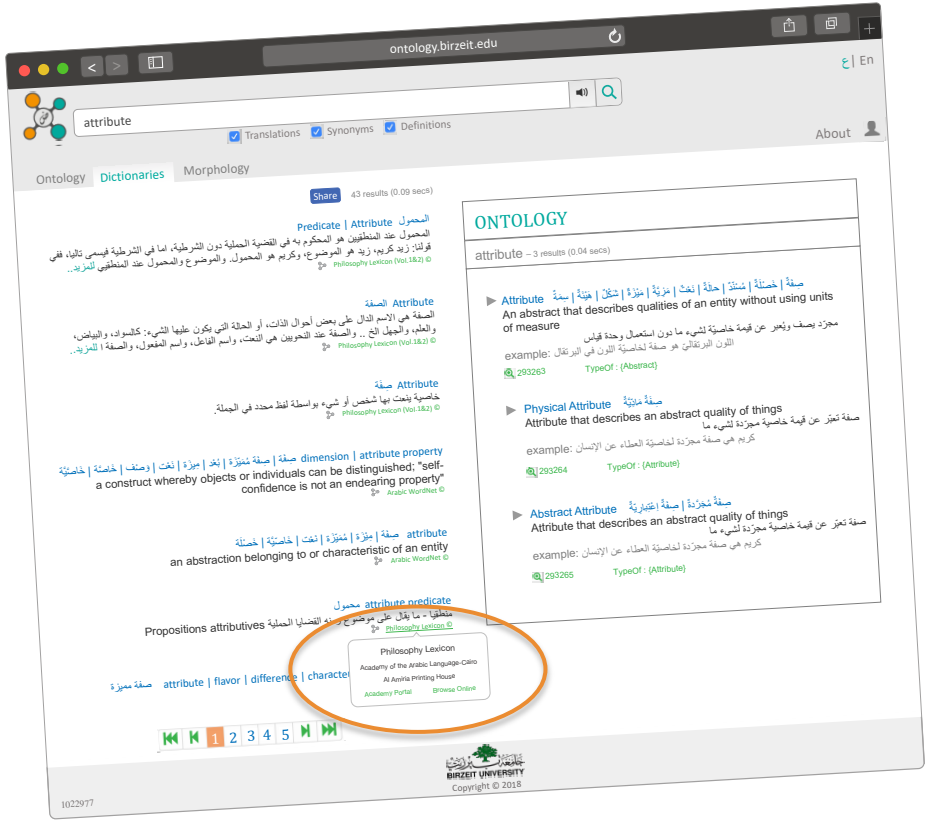
- طباعة يدوية، وإعادة صياغة لبنية المعلومات، وتخزينها في قاعدة بيانات محوسبة.
- عمل طوعي، ولم يتم تمويله من أية جهة داعمة.
- عمل غير ربحي.



https://ontology.birzeit.edu

حقوق الملكية

- تم التواصل مع أصحاب المعاجم وأخذ إذنهم.
- **ترويج للمعاجم**، وإظهار معلومات المعجم بجانب كل تعريف (المؤلف، الناشر، الصفحة، ...)



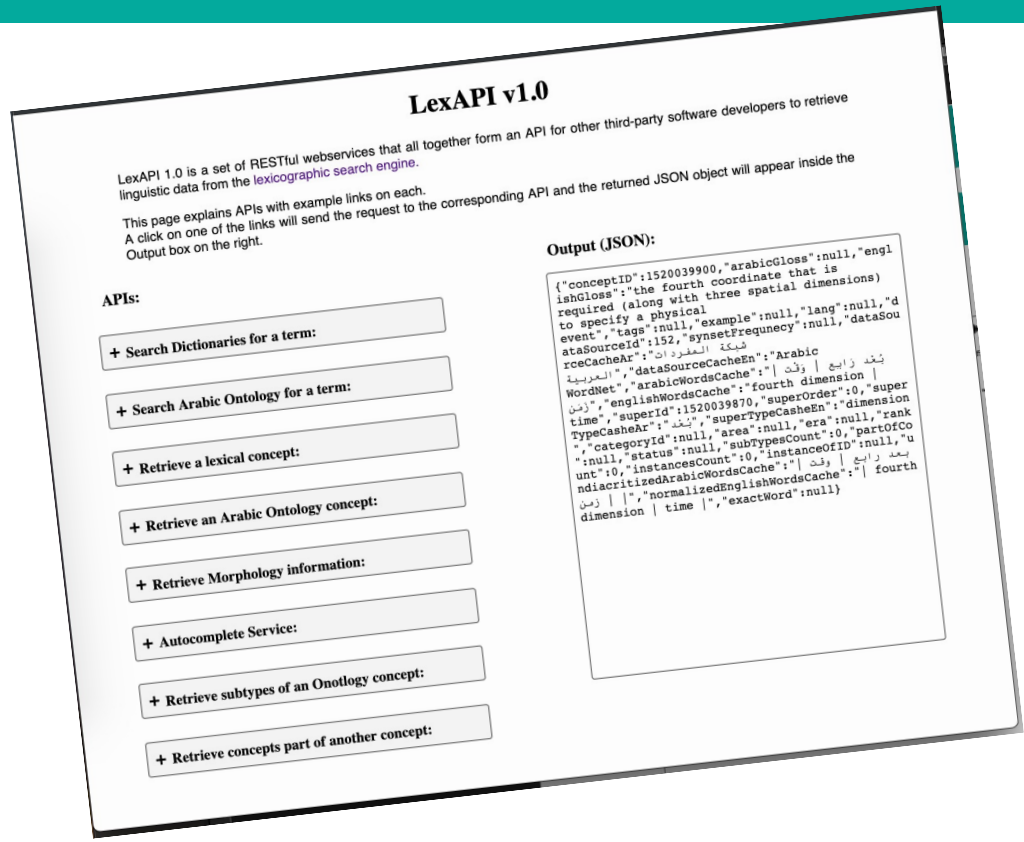
https://ontology.birzeit.edu

محرك بحث معجمي

LexAPI

RESTful web services to directly retrieve linguistic data from the lexicographic database (ontology, Morph solutions, synonyms, translations, definitions, etc.).

- Request API Token



الأنطولوجيا العربية
Arabic Ontology

(Formal Arabic Wordnet)

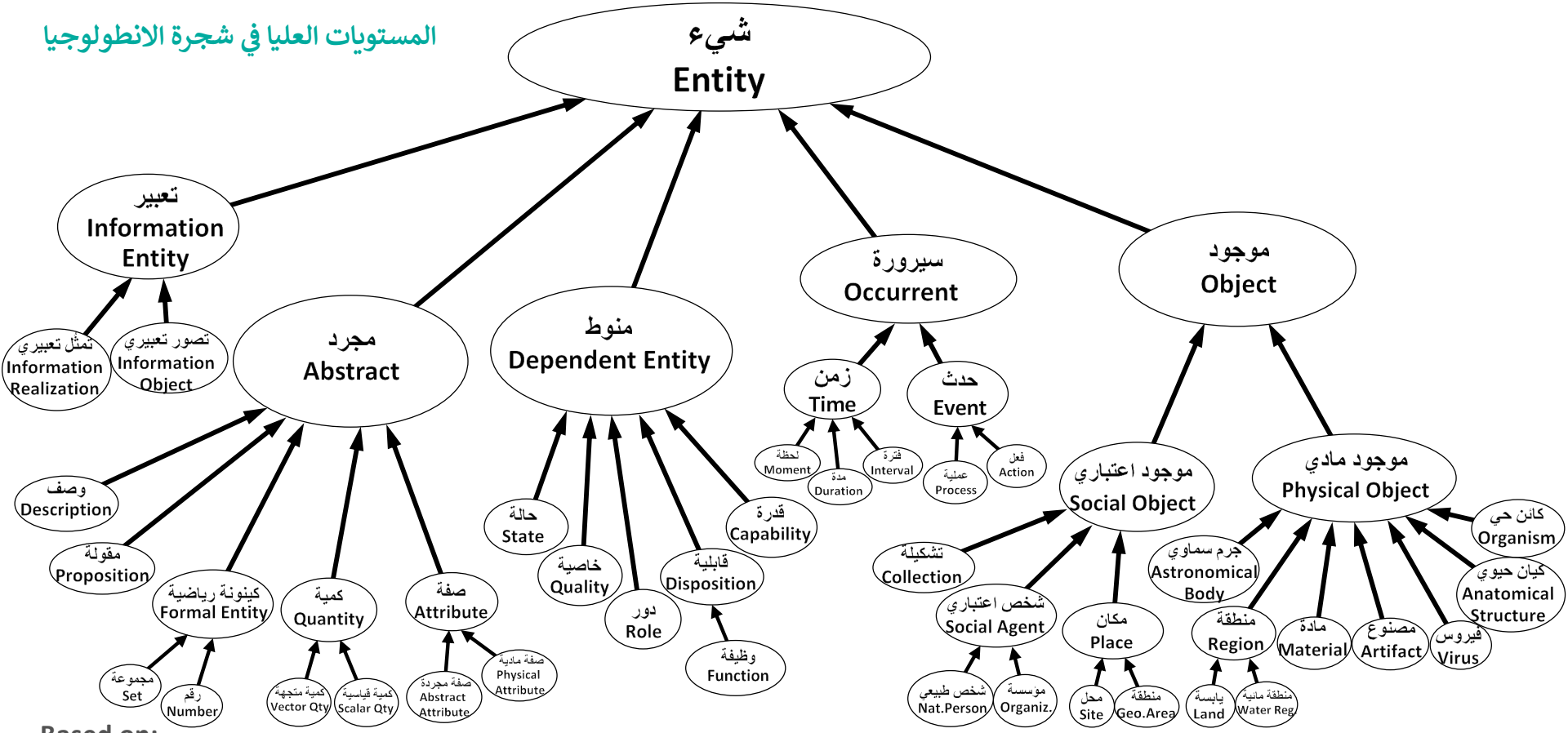
الأنطولوجيا العربية

- تستعمل كـ(formal Arabic wordnet)
- عمل حاسوبي وفلسفي ولغوي.
- الانطولوجيا العربية: وصف معاني الكلمات اللغة العربية، وعلاقات بين هذه المعاني.
- تصنيف مفاهيم (معاني) الكلمات العربية دلالياً، وتمثيلها بلغة المنطق ليستطيع الحاسوب فهمها.
- تختلف عن المعاجم كون الحكم على صحة المعلومات فيها يعتمد على ما توصلت إليه العلوم، وليس على ما شاع بين الناس.
- مربوطة بشبكة المفردات الإنجليزية (وردنت) والويكي بيانات.

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

الأنطولوجيا العربية

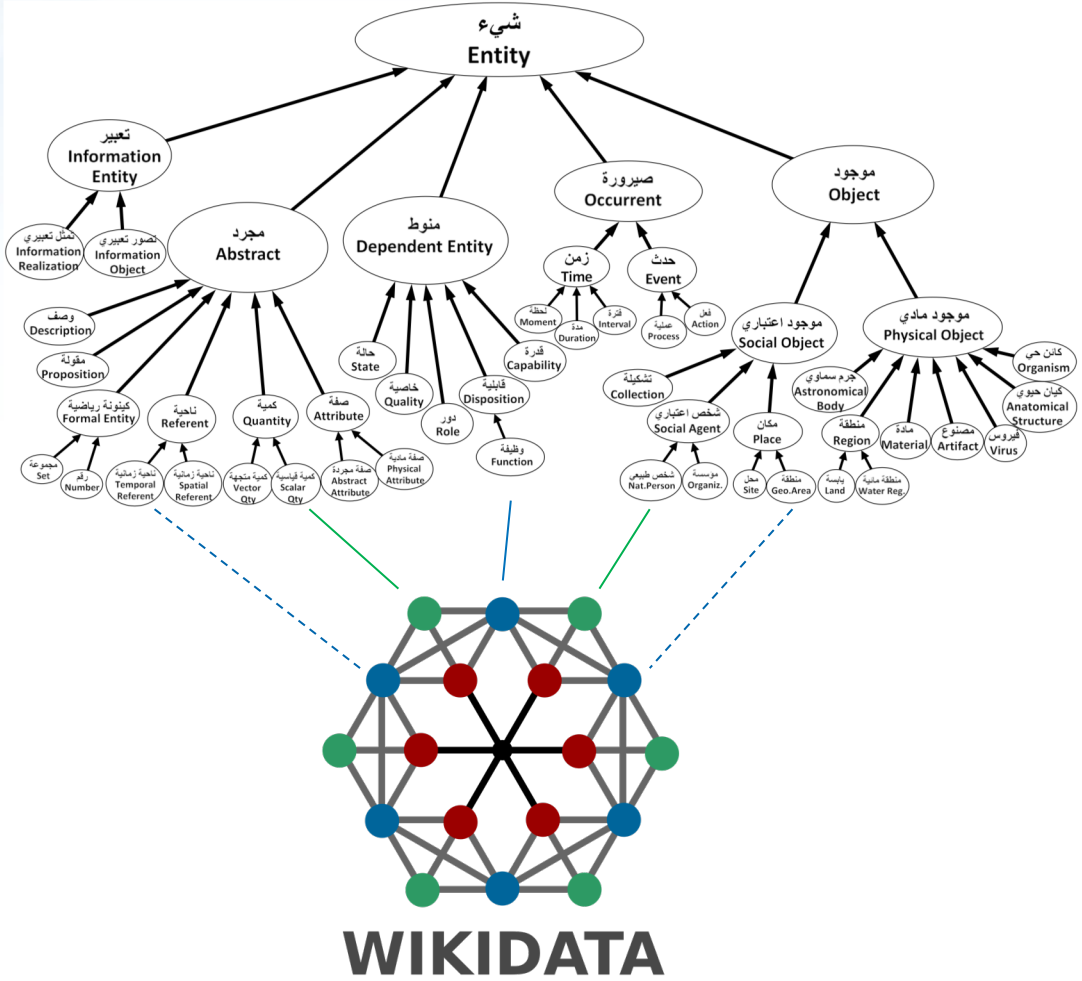
المستويات العليا في شجرة الانطولوجيا



Based on:

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

ربط الأنطولوجيا بالويكيبيديا



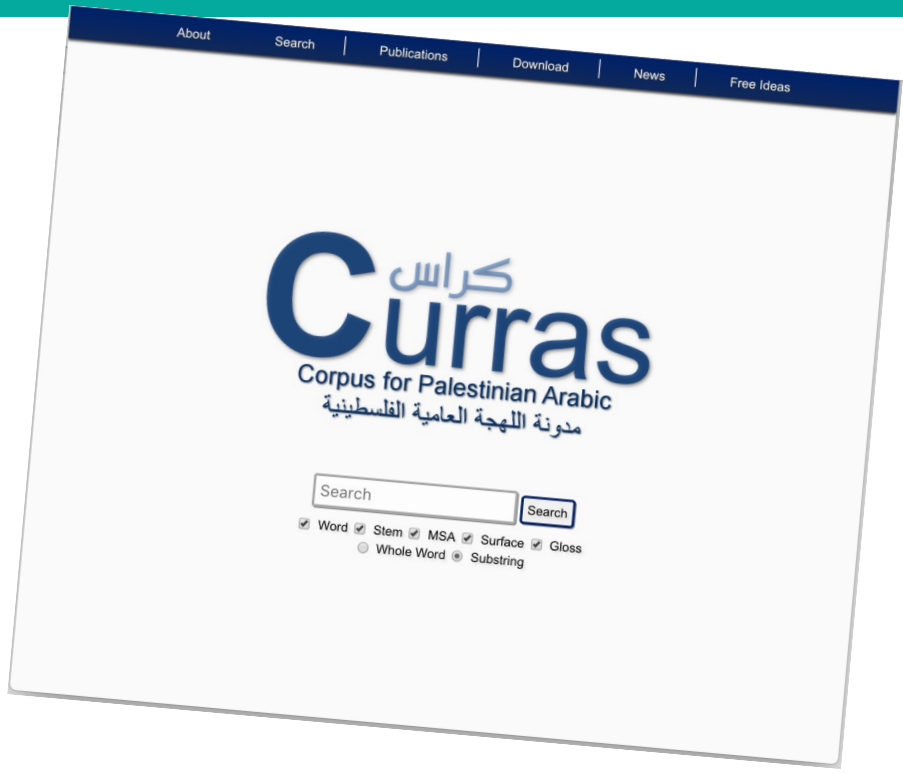
- تم ربط كل مفهوم بالأنطولوجيا بمقابله بشبكة بيانات الويكيبيديا (ويكي بيانات).
- وبالتالي، إنشاء شبكة معرفية (**Knowledge Graph**) مؤصلة بالمفاهيم العربية.
- تطبيقات ذلك مثل، إجابة عن أسئلة معرفية عامة بالعربية.

مدونة اللهجة العامية

Dialectal Corpus

مدونة اللهجة العامية

- عدد ضخم من النصوص المكتوبة باللهجة العامية.
- تصريف كل كلمة بالمدونة ووسمها بحوالي 16 سمة (سوابق، لواحق، جذع الكلمة، فرع الكلمة، المقابل بالفصحى، المقابل بالإنجليزية، ...).
- تحتوي 60 ألف كلمة.



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

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

مدونة اللهجة العامية

تصريف كل كلمة بالمدونة ووسمها بحوالي 16 صفة (سوابق، لواحق، جذع الكلمة، فرع الكلمة، المقابل بالفصحى، المقابل بالإنجليزية، ...).

شو القصة ليش مأزمة؟

60k ...

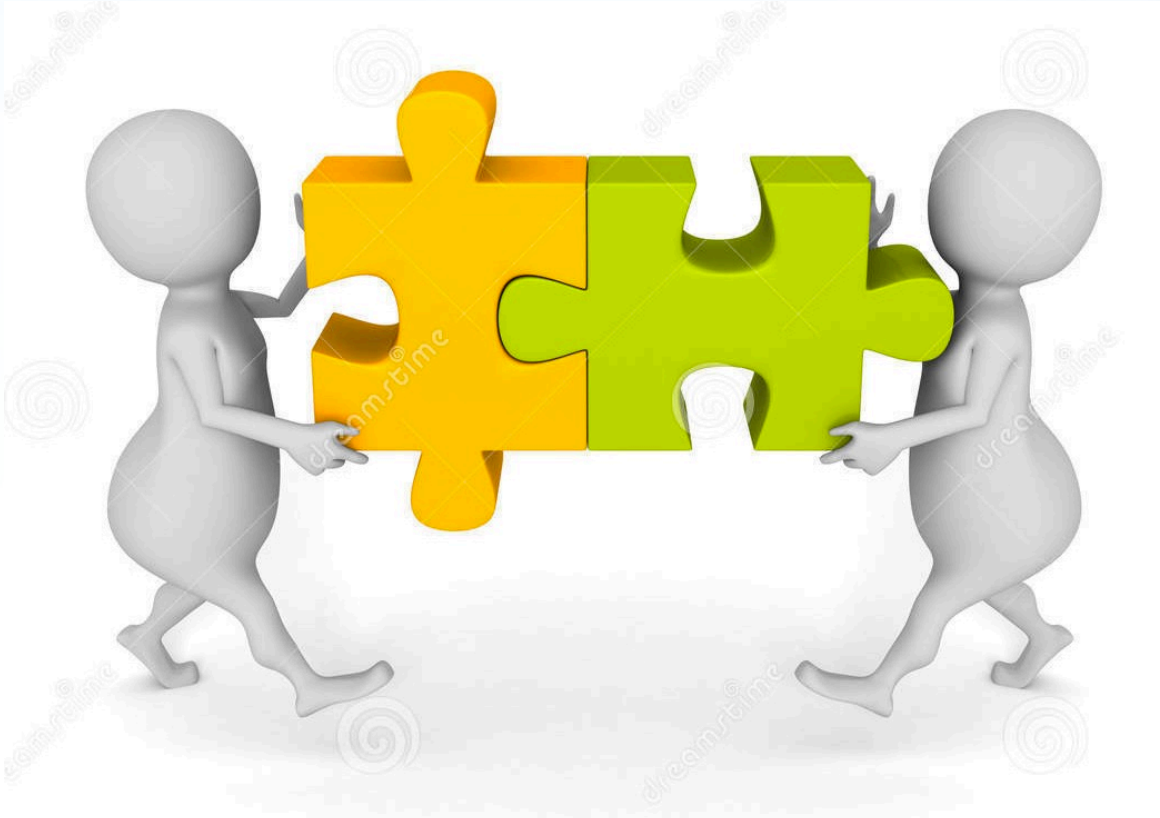
Lemma: قصة١
POS: ال/DET +قص/NOUN + ة/NSUFF_FEM_SG
Person: N/A
Gender: female
Gloss: story;stories_
....

Lemma: شو١
MSA: ماذا١
POS: شو/INTERROG_PRON
Person: N/A
Gender: male
Gloss: what;which
....

مدونة اللهجة العامية

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

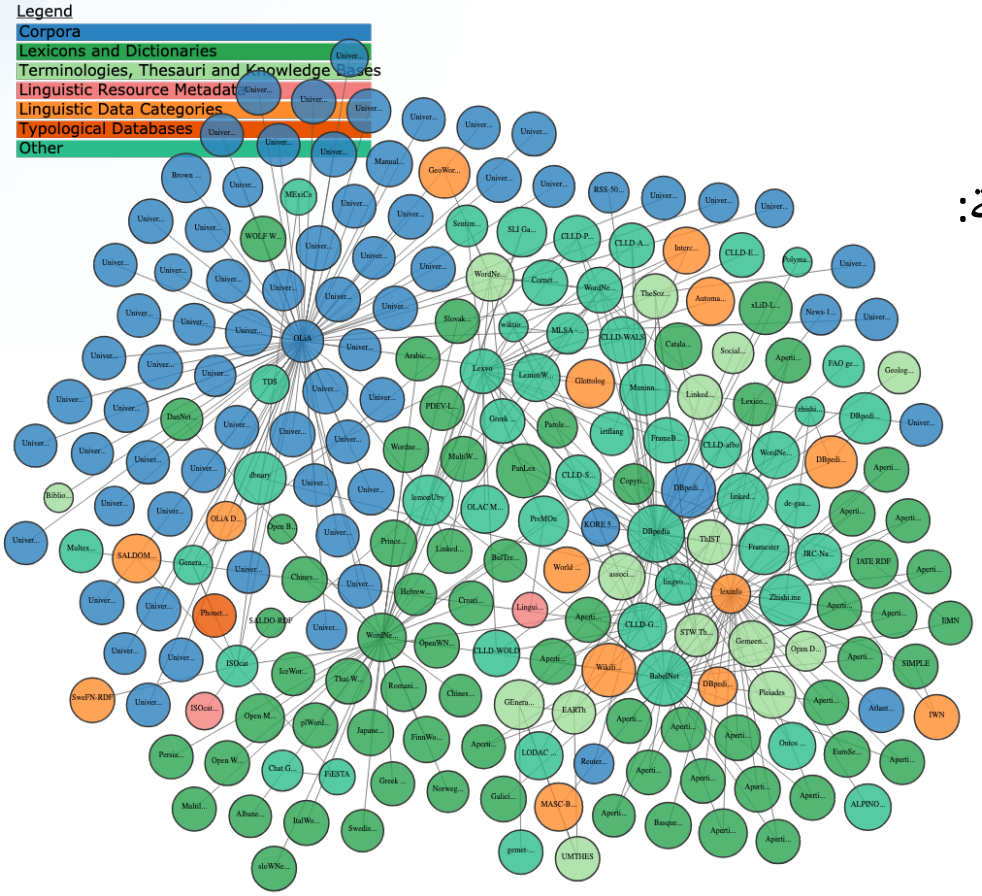
بنتنا زي كعك العيد بتنفق بسرعة. العريس : والله انتو عيلة ما بشتريكوا بقمعة سيجارة , اخر اشي بقولكوا اياه غروس مالبورو شو قلتوا ? الام : اسمع ولا , خاي ا	M	1	Verb	said;say;be_said;(for_example)_[CALIMA]	b/PROG_PART+A/IV1S+qwl/IV+kw/IVSUFF_DO:2P	باقولكو	قال 1	قول	قال 1	بقولكوا
زبون يا بقتل امي . منال : 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/IV3MS+qwl/IV+k/IVSUFF_DO:2MS	بيقولك	قال 1	قول	قال 1	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 شوه الأز	M	3	Verb	said;say;be_said;(for_example)_[CALIMA]	b/PROG_PART+y/IV3MS+qwl/IV+k/IVSUFF_DO:2MS	بيقولك	قال 1	قول	قال 1	بقولك
مة شو القصة ليش مأزمة ? بقولك يوم السبت عادة هيك , بتطلع الاثنين 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/IV3MS+qwl/IV+k/IVSUFF_DO:2MS	بيقولك	قال 1	قول	قال 1	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/IV3MS+qwl/IV+k/IVSUFF_DO:2MS	بيقولك	قال 1	قول	قال 1	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/IV3MS+qwl/IV+k/IVSUFF_DO:2MS	بيقولك	قال 1	قول	قال 1	بقولك



ربط بين المصادر اللغوية

شبكة عالمية من البيانات المعجمية المترابطة

The Linguistic Linked Open Data Cloud (LLODC)



التوجه العالمي الحديث للتعامل مع المصادر اللغوية:

- جهود بحثية مشتركة لإنشاء شبكة عالمية من البيانات المعجمية المترابطة.
- يتم تمثيل كل مدخلة ومفهوم وتعريف... في كل مصدر لغوي بطريقة (lemon) وربطها مع مقابلاتها في المصادر واللغات الأخرى.

قيد البحث والتطوير حالياً

بناء **شبكة بيانات لغوية ضخمة** للغة العربية تحتوي جميع المستويات التصريفية والاشتقاقية والدلالية والعامية.

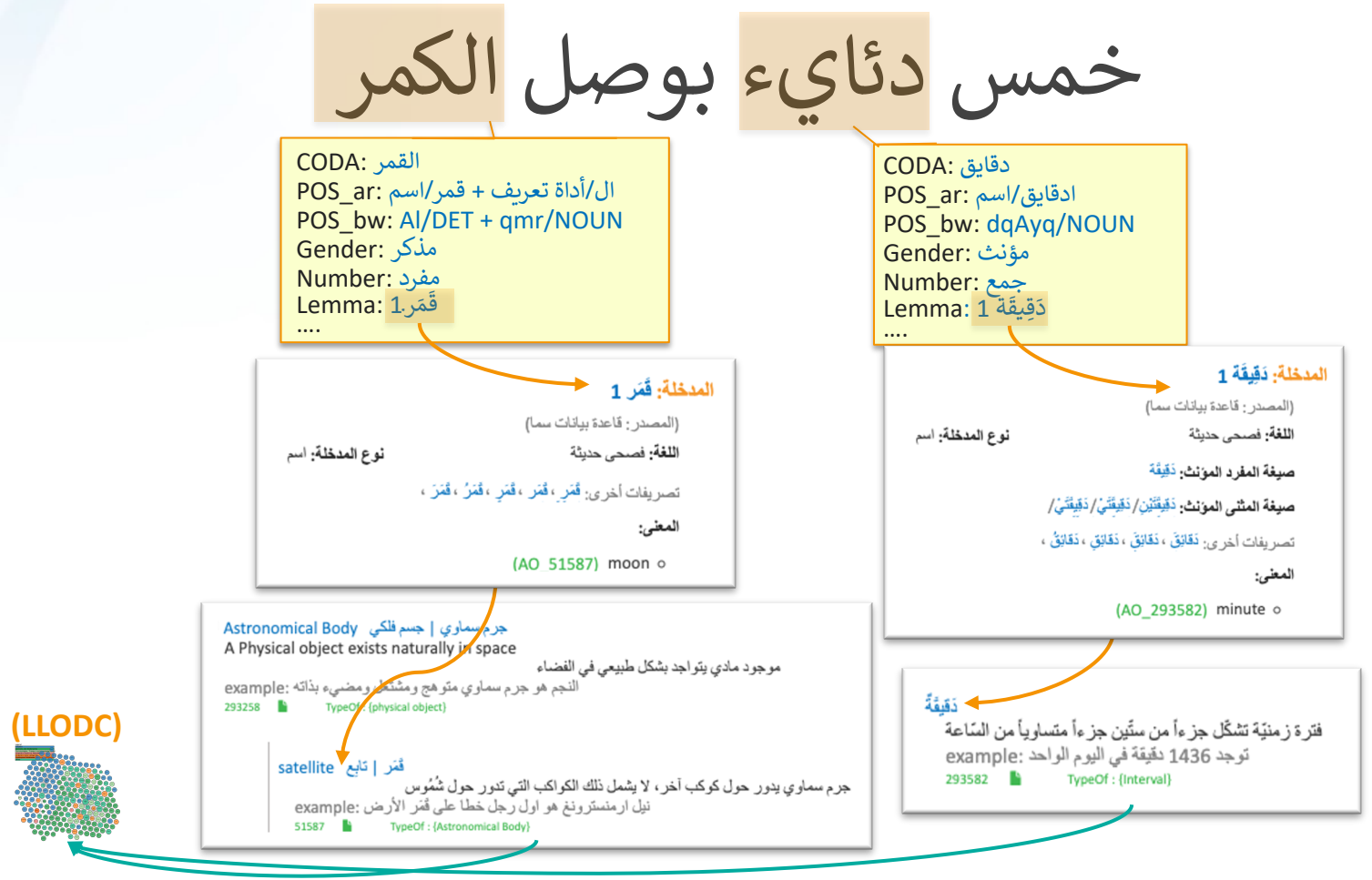
Arabic Lemma Index

قائمة بجميع المدخلات العربية القديمة والحديثة والفصحى والعامية، وتحديد سماتها الصرفية ومعانيها وعلاقاتها الاشتقاقية.

Arabic Linked Data Cloud

ربط المعاجم من خلال ربط كل فرع (lemma) وتصريفاتها (inflections) لكل مدخلة في الـ150 معجماً، ثم ربطها جميعاً بشبكات المصادر اللغوية الأجنبية.

فهم وربط لجميع المستويات اللغوية



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

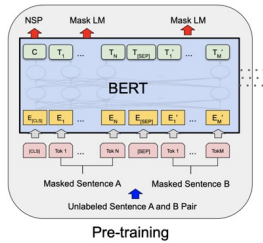
❖ Arabic context-gloss pairs Dataset (167k)

- 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)	[CLS] قصيدة من عيون الشعر [SEP] عين الشيء : نفسه ، ذاته	False
[SEP] (تستعمل للتوكيد)	[CLS] جاء القوم أعينهم [SEP] عين الشيء : نفسه ، ذاته	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

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

شكرا لاهتمامكم

مصطفى جرار – جامعة بيرزيت
mjarrar@birzeit.edu