Natural Language Processing

Morphology

Mustafa Jarrar

Birzeit University

Watch this lecture and download the slides

Course Page: http://www.jarrar.info/courses/NLP/
More Online Courses at: http://www.jarrar.info

Acknowledgement:
This lecture is based on (but not limited to) to the lecture notes found in [……]

Natural Language Processing

Morphology

In this lecture:

- Part 1: **What is morphology**
- Part 2: Lexeme and Lemma
- Part 3: Stem and affixes
- Part 4: Part-of-Speech

What is Morphology
التصريف

The study of how words are formed

Words are made up of smaller meaning-bearing units (morphemes).

Morpheme: smallest meaningful unit of a word. Morphemes can be: stems, suffixes, prefixes …)

المدرسين
سيكتبها

Books
Unliked

Morphology is the identification, analysis, and description of the morphemes of a given word, including its root, part-of-speech, gender, number, etc.

Types of Morphology : Inflectional Morphology and derivational morphology

Inflectional Morphology (تصريف)

Modifying words to express **different grammatical categories** (gender, number, voice, mood, aspect, case, etc.)
forming new wordforms without changing its core meaning.

Verbs

كتب

كتبت، كتبها، سأكتبها، يكتب، أكتب، يكتبون، سيكتبها، وسيكتبها

....

Write

writes, wrote, writing

Nouns

عالم

عالمة، عالمين، عالمتين، علماء، عالمات وعلمائنا بعلمائكما

...

Scientist

Scientists

Inflectional Morphology will be the focus of this this lecture

Derivational Morphology (اشتقاق)

Modifying words to express **different meaning and/or part-of-speech** forming a new lemma (i.e., drive lemma from another lemma)

كتب	write
كتابة	writing
كاتب	writer
مكتوب	weakness
مكتبة	weaken
مكتب	reddish
كتيّب	relational
كاتبة	relationally
	testify
	eatable
	guidance

English Derivational Morphology (الاشتقاق بالإنجليزية)

Examples of derivational patterns in English

verb-to-noun (agent):	er	**write → writer**
verb-to-adjective:	able	**eat → eatable**
verb-to-noun:	ance	**guide → guidance**
noun-to-adjective:	al	**relation → relational**
noun-to-verb:	fy	**test → testify**
adjective-to-noun:	ness	**weak → weakness**
adjective-to-verb:	en	**weak → weaken**
adjective-to-adjective:	ish	**red → reddish**
adjective-to-adverb:	ly	**relation → relationally**
….	…	**….**

Arabic Derivational Morphology (الاشتقاق بالعربية)

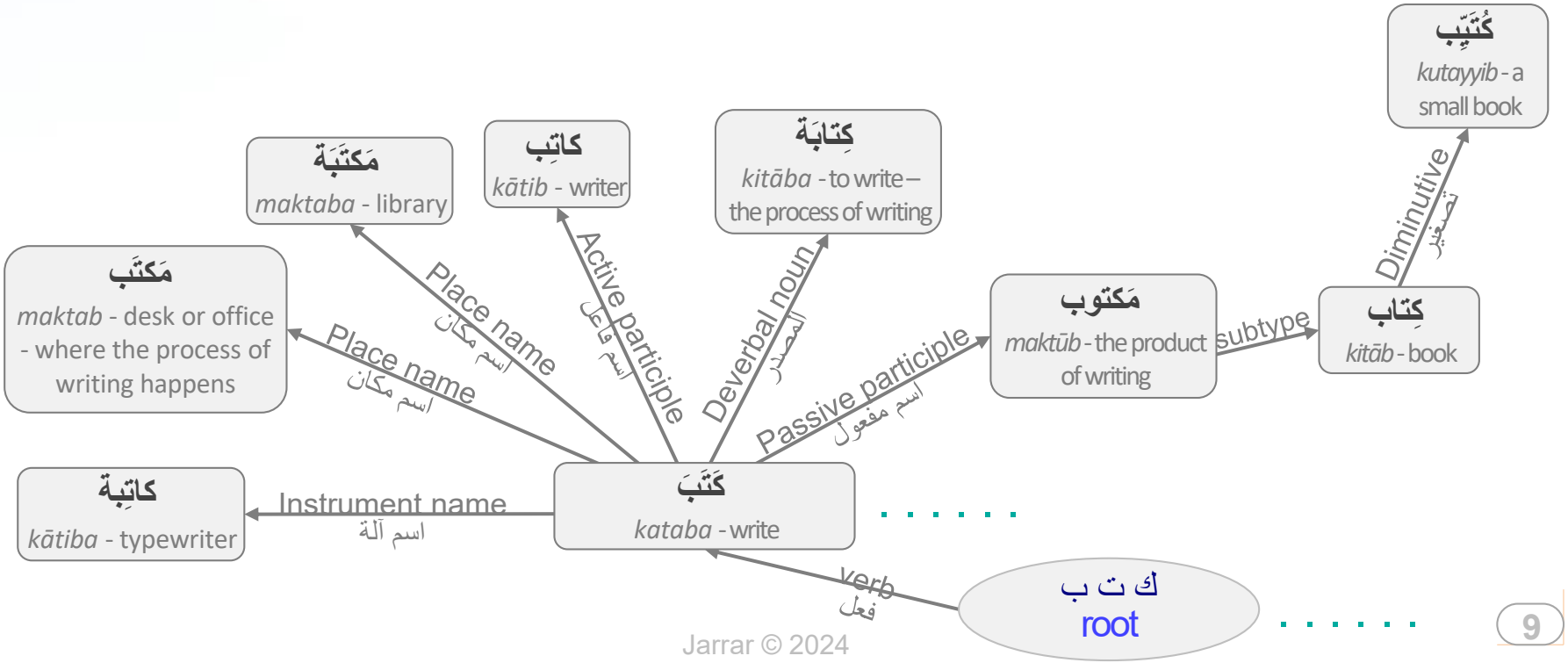
Derivational morpology in Arabic is

- mostly templatic and follows derivational patters (أوزان اشتقاقية).
- changes the meaning (morpho-semantic category) rather than part-of-speech
- Examples:

كتب	فعل
كتابة	مصدر أصلي
كاتب	اسم فاعل
كتَّيب	مبالغة اسم فاعل
مكتوب	اسم مفعول
مكتبة	اسم مكان
مكتب	اسم مكان
كتيّب	تصغير
كاتبة	اسم آلة
.....	...

Arabic Derivational Morphology (الاشتقاق بالعربية)

- Derivational morpology in Arabic can be seen as **morpho-semanitc relationships** between lemmas (some are through a sense), as suggested in (Jarrar, 2021).
- A large graph of Arabic morpho-semanitc relationships is available at (Jarrar & Amayreh, 2019).
- Example (simplified, some relations should be rough a sense):



Natural Language Processing

Morphology

In this lecture:

- **Part 1:** What is Morphology
- **Part 2:** **Lexeme and Lemma**
- **Part 3:** Stem and affixes
- **Part 4:** Part-of-Speech

Lexeme (لكسيم/وحدة صرفية)

Lexeme: lexical meaning for a **set of words** that related through inflections

Verb Lexemes:

{write, writes, wrote, writing}

{كتب، كتبت، يكتب، أكتب، كتبها، سأكتبها، يكتب، يكتبون، وسيكتبها ...}

Noun Lexemes:

{scientist, scientists}

{عالم، عالمة، عالمين، عالمتين، علماء، عالمات، وعلمائنا، ...}

Lemma (مدخلة معجمية)

Lemma: is the dictionary form (also called canonical form and citation form) of a set of inflections (i.e., to represent a lexeme).

One of the words in the lexeme is selected **conventionally/اصطلاحا** to refer to a lexeme.

Verb lemmas: typcially the past singular misculin 3rd form (الماضي المفرد المذكر الغائب)

{**write**, writes, wrote, writing}

{**كتب**، كتبت، يكتب، أكتب، كتبها، سأكتبها، يكتب، يكتبون، وسيكتبها ...}

Noun Lemmas: typcially the singular form (المفرد)

{**scientist**, scientists}

{**عالم**، عالمة، عالمين، عالمتين، علماء، عالمات، وعلمائنا، ...}

Lemma (مدخلة معجمية)

Issues about lemmas

Since lemmas are chosen (**conventionally/اصطلاحا**) lexicons do not necessarily agree on them:

- Some lexicons split lemmas based on the common-use (شيوع الاستخدام)
  - **include feminine and masculine forms** as separate lemmas, like (أشهب - شهباء) (قاصر - قاصرة) ... if all are commonly used
  - **feminine forms if no masculine** like (جامعة, تفاحة)
  - **Plural or dual forms** (تخوم، تسعينيات، مُذْهَبات) or (ثقلان، اثنان، أخبثان) if the singular forms is not common.
- Some lexicons **split lemmas if they are semantically very different**, like in SAMA (جامعة-1, university) and (جامعة-2, league) while Ghani consider it one.

Lemmatization Examples

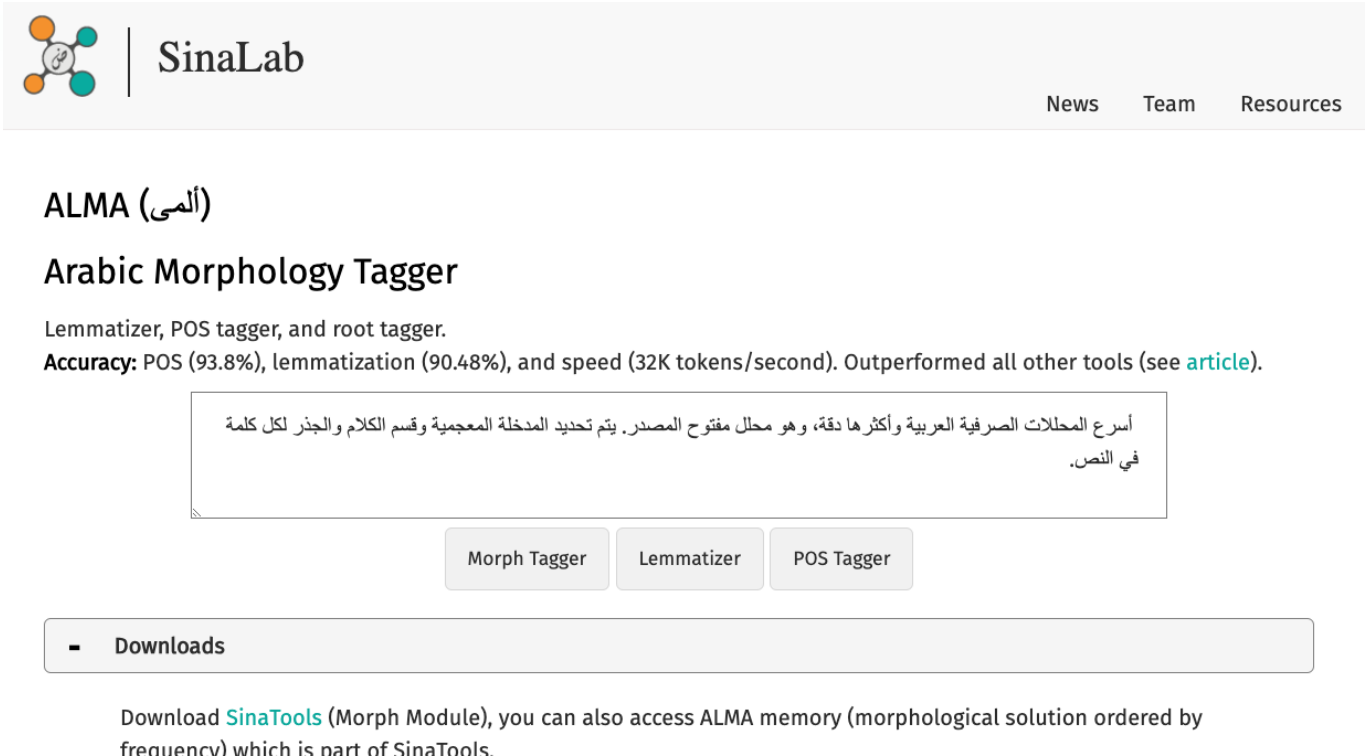
Word	Lemma
وسيكتبها	كَتَبَ
إكتبلي	كَتَبَ
مكتباتهم	مَكْتَبَةٌ
مكاتبكم	مَكْتَبٌ
بمكتباتكن	مَكْتَبَةٌ
فكتابها	كِتَابٌ
كِتابي	كِتَابٌ
كُتَيِّبنا	كُتَيِّبٌ
فكِتاباتنا	كِتابةٌ
بكَتَبَة	كَاتِبٌ
وكُتّابهم	كَاتِبٌ
كُتُب	كِتَابٌ

Word	Lemma
أبيات	بَيْتٌ2
بيوت	بَيْتٌ2 بَيْتٌ1
هذا	هَذَا
هذه	هَذَا
هؤلاء	هَذَا
بكم	بِ2
لكم	لِ2
على	عَلَى2
عليكن	عَلَى2

Lemmatization

Lemmatizer: a program that returns the lemma of a given word/sentence

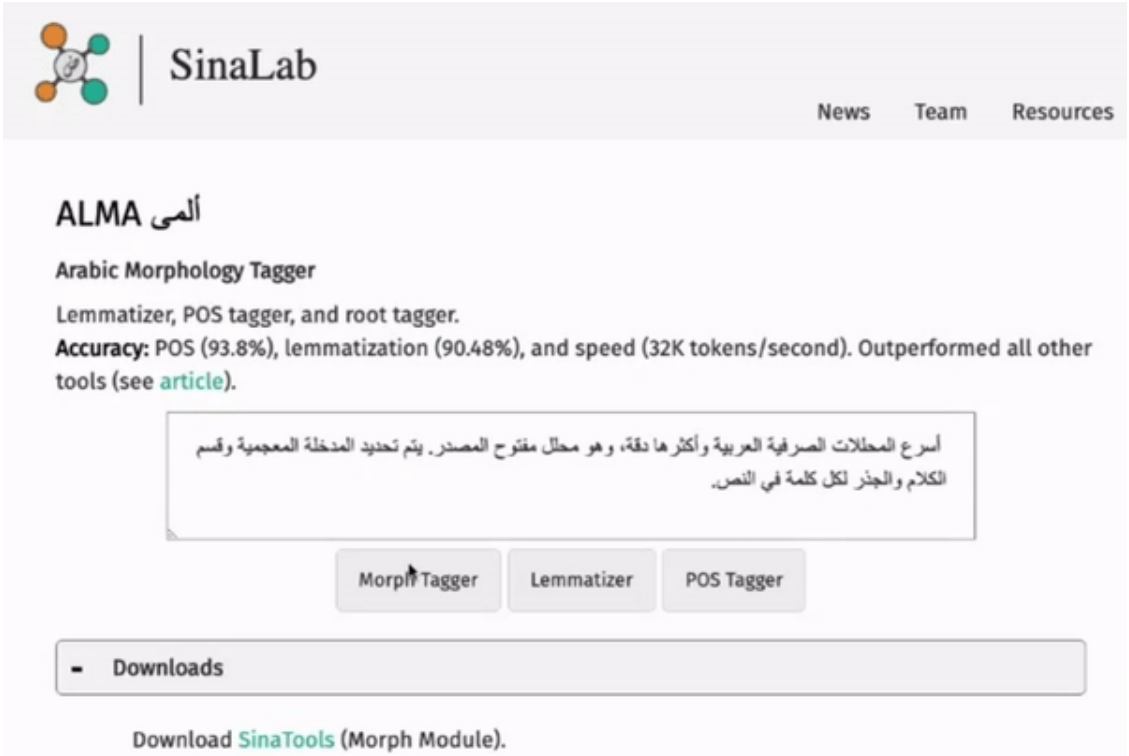
https://sina.birzeit.edu/alma



Lemmatization

Lemmatizer: a program that returns the lemma of a given word/sentence

https://sina.birzeit.edu/alma



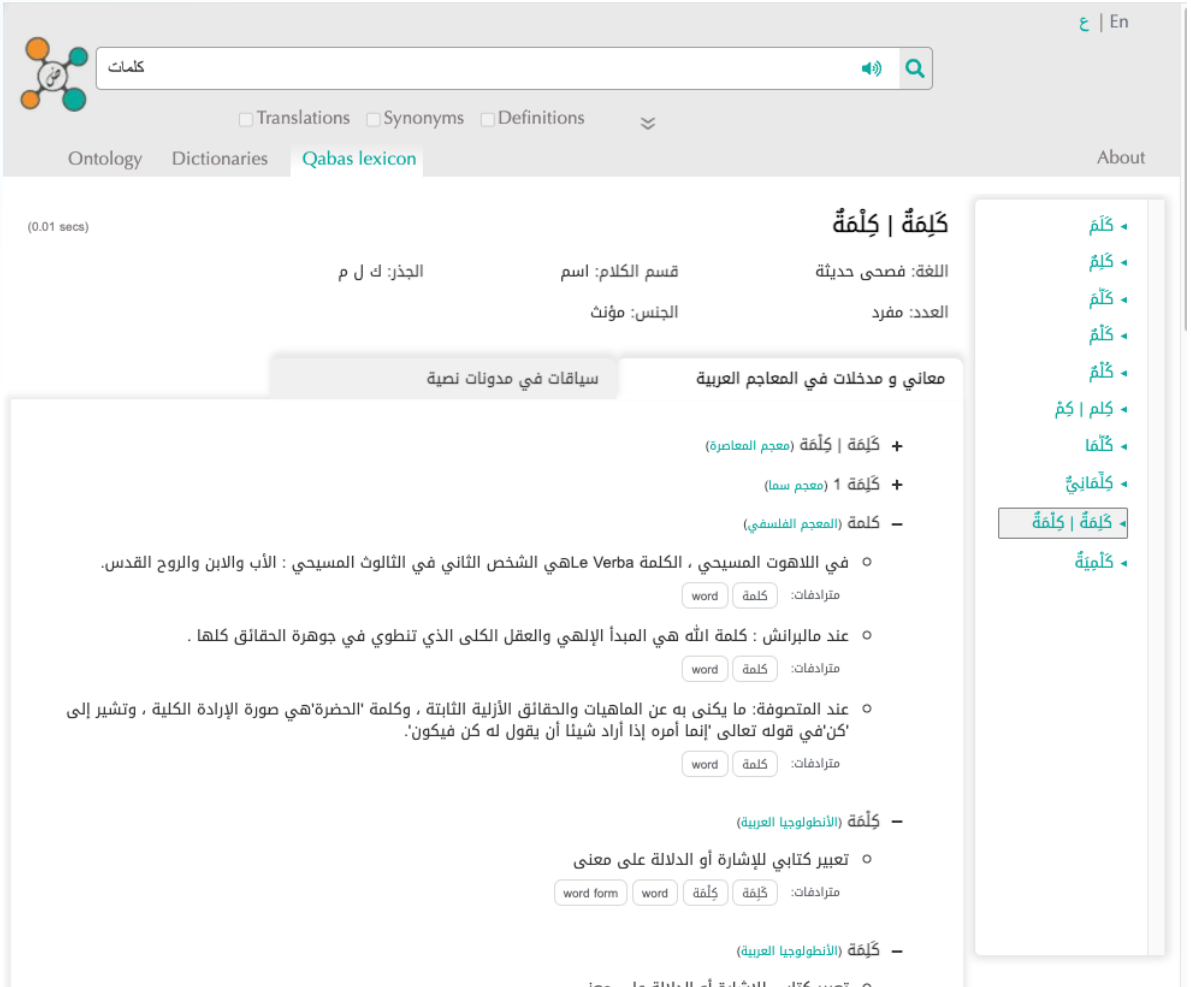
Qabas

Open-source Lexicographic Database

https://sina.birzeit.edu/qabas

Qabas

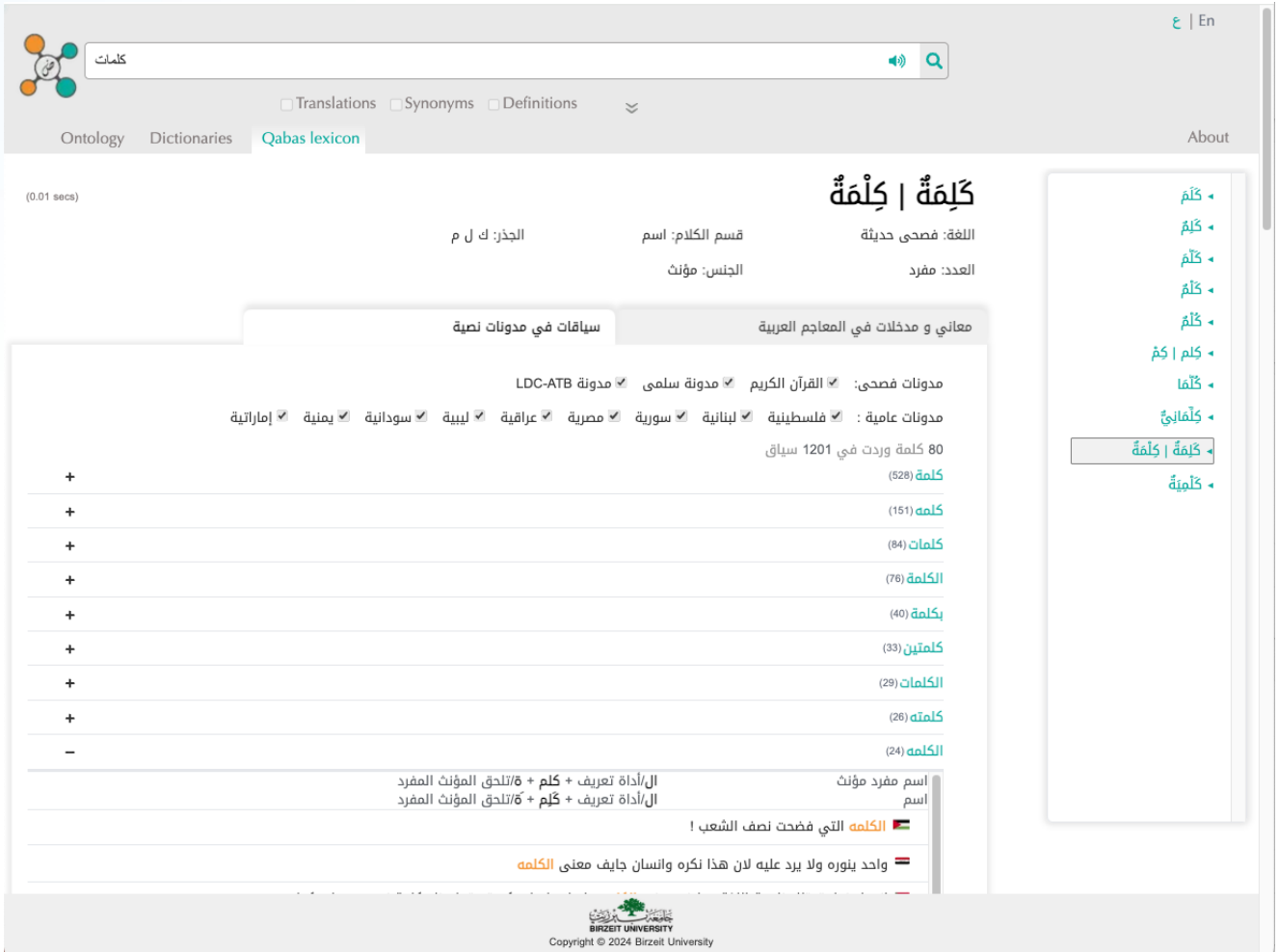
Open-source Lexicographic Database



https://sina.birzeit.edu/qabas

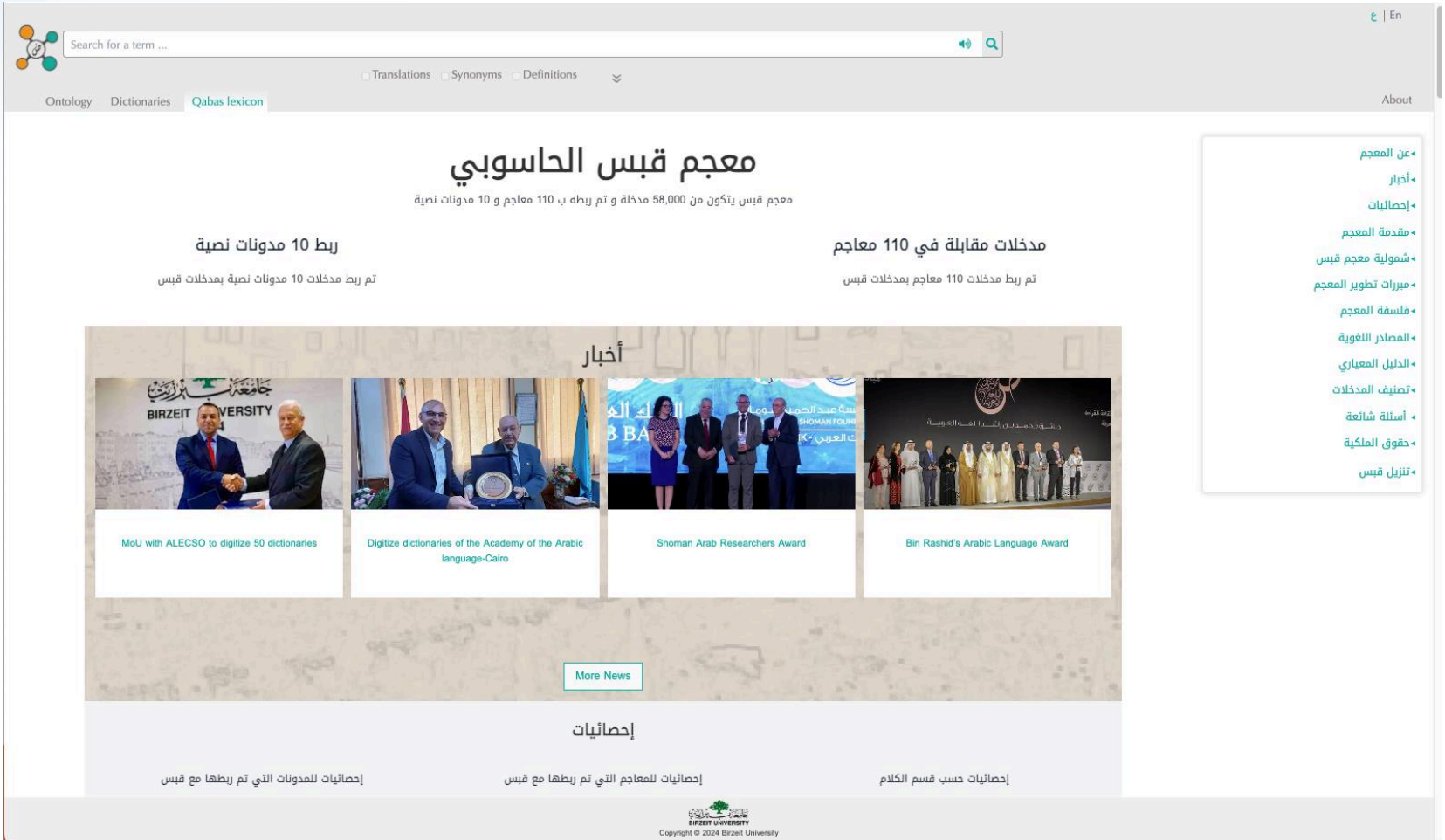
Qabas

Open-source Lexicographic Database



Qabas

Open-source Lexicographic Database



Natural Language Processing

Morphology

In this lecture:

- ❑ Part 1: What is Morphology
- ❑ Part 2: Lexeme and Lemma
- ❑ Part 3: **Stem and Affixes**
- ❑ Part 4: Part-of-Speech

Root and Stem

Root (جذر): is the primary lexical unit of word, and which **carries aspects of semantic content**.

- Roots in Arabic are **decided by linguists** (who sometimes disagree).
- Some words may have more than one root, سنة (س ن ي)(س ن و)

Stem (ساق): is the part that is common to all its inflected variants (no agreed definition), thus it **depends on the stemming algorithm**.

Roots and Stems in English are most likely the same, but in Arabic Roots and Stems are not the same.

Examples:

Word	Lemma	Root	Stem
سيكتبها	كَتَبَ	ك ت ب	كتب
مكتباتهم	مَكْتَبَةٌ	ك ت ب	مكتب
مكاتبكم	مَكْتَبٌ	ك ت ب	مكاتب
كتابي	كِتَابٌ	ك ت ب	كتاب

Inflectional Affixes (زوائد صرفية)

Prefixes (سوابق): attached before the stem to form a word form.

يكتب، سيكتب، وسيكتب، الكتاب، فكتب، أكتب، كالكتاب،

Suffixes (لواحق): attached after the stem to form a word form.

book**s**, book**ed**, happ**iness**, usual**ly**, help**ful**, …
كتبت، كتبها، كتبلك، كتبكم، كتبكما، كتبهن، ....

Infixes (أواسط): inserted inside the stem to form a word form.
اجتهد، ابتاع، ...

Stemmer: a program that returns the stem of a given word, after removing affixes.

Prefixes (سوابق)

attached after the stem to form a word form

يكتب، سيكتب، وسيكتب، الكتاب، فكتب، أكتب، كالكتاب،

In English:

- There are **no inflectional prefixes**;
- English uses suffixes to generate inflectional forms.
- Prefixes are derivational prefixes thus these words are new lemmas, not-inflections, such as (unhappy, asymmetric, demotivate, bilingual, cooperation, disappear, illegal…)

In Arabic:

Very complex and rich - 174 prefix types (115 un diacritized) in SAMA, such as:

أ، ا، ال، ب، بال، ف، فأ، فا، فال، فب، فبال، فك، فكال، فل، فلأ، فلا، فلاأ، فلاا، فلان، فلات، فلاي، فلاآ، فلل، فلن، فلت، فلي، فلآ، فماأ، فماا، فمان، فمات، فماي، فماآ، فن، فسأ، فسا، فسن، فست، فسي، فسآ، فت، في، فآ، ك، كال، ل، لأ، لا، لأأ، لاا، لان، لات، لاي، لاآ، لل، لن، لت، لي، لآ، ماأ، ماا، مان، مات، ماي، ماآ، ن، سأ، سا، سن، ست، سي، سآ، ت، و، وأ، وا، وال، وب، وبال، وف، وك، وكال، ول، ولأ، ولا، ولاأ، ولاا، ولان، ولات، ولاي، ولاآ، ولل، ولن، ولت، ولي، ولآ، وماأ، وماا، ومان، ومات، وماي، وماآ، ون، وسأ، وسا، وسن، وست، وسي، وسآ، وت، وي، وآ، ي، آ، .....

Suffixes (لواحق)

attached before the stem to form a word form.

book**s**, book**ed**, happ**iness**, usual**ly**, help**ful**, …

كتبت، كتبها، كتبلك، كتبكم، كتبكما، كتبهن، ....

In English:

- changes the grammatical properties of a word within its syntactic category
- 8 suffix types: Verbs{make**s**, book**ed**, eat**ing**, eat**en**}, Nouns{girl**s**, ox**en**}, adjectives and Adverbs {larg**er**, larg**est**}.
- Irregulars: sheep, children, ran, drunk, ….

In Arabic:

Very complex and rich - 416 suffix types (208 un diacritized) in SAMA, such as:

يهن،يهما،يهم،يها،يه،يني،ينهن،ينهما،ينهم،ينها،ينه،ينني،يننا،ين،يكن،يكما،يكم،يك،ي،و هن،و هما،و هم،و ها،وه،وني،ونهن،ونهما،ونهم،ونها،ونه،وننني
،وننا،ونكن،ونكما،ونكم،ونك،ونا،ون،وكن،وكما،وكم،وك،وا،و ،هن،هما،هم،ها،ه،ني،نهن،نهما،نهم،نها،نه،ننني،ننا،نكن،نكما،نكم،نك،ناهن،ناهما،ناهم،ناه
ا،ناه،ناكن،ناكما،ناكم،ناك،نا،ن،مو هن،مو هما،مو هم،مو ها،موه،موني،مونا،ماهن،ماهما،ماهم،ماها،ماه،ماني،مانا،ما،م،كن،كما،كم،ك،تيهن،تيهما،تيهم،ت
يها،تيه،تينا،تين،تيكن،تيكما،تيكم،تيك،تي،تهن،تهما،تهم،تها،ته،تني،تنهن،تنهما،تنهم،تنها،تنه،تنني،تننا،تنا،تن،تمو هن،تمو هما،تمو هم،تمو ها،تموه،تموني،
تمونا،تماهن،تماهما،تماهم،تماها،تماه،تماني،تمانا،تما،تم،تكن،تكما،تكم،تك،تاي،تاهن،تاهما،تاهم،تاها،تاه،تاني،تانا،تان،تاكن،تاكما،تاكم،تاك،تا،ت،ة،اً،ا
ي،اهن،اهما،اهم،اها،اه،اني،انهن،انهما،انهم،انها،انه،انني،اننا،انكن،انكما،انكم،انك،انا،ان،اكن،اكما،اكم،اك،اتي،اتهن،اتهما،اتهم،اتها،اته،اتنا،اتكن،اتكما،
اتكم،اتك،ات،ا،آتي،آتهن،آتهما،آتهم،آتها،آته،آتنا،آتكن،آتكما،آتكم،آتك،آت، …

Natural Language Processing

Morphology

In this lecture:

- ❑ **Part 1:** What is Morphology
- ❑ **Part 2:** Lexeme and Lemma
- ❑ **Part 3:** Stem and affixes
- ❑ **Part 4:** **Part-of-Speech (Short Overview)**

Part-of-Speech (اقسام الكلام)

Also called **word class**, **lexical class**, and **lexical category**

a category of words that have similar grammatical properties.

Words have the same POS have similar syntaxic behavior, and sometimes similar morphology

In English:

- **noun** (e.g., book)
- **verb** (e.g., booked)
- **adjective** (e.g.,)
- **adverb** (e.g., very, quite)
- **pronoun** (e.g., he, them)
- **preposition** (e.g., in, of)
- **conjunction** (e.g., and, but)
- **interjection** (e.g., ops, alas)
- **numeral** (e.g., one, two)
- **article** (e.g., the, a , an)

In Arabic:

- **noun**
- **verb (PV, IV, CV)**
- **Letter/Functional words**
  words that are not nouns or verbs)

Next Lecture

➔Part-of-Speech Tagging

References

Mustafa Jarrar, Tymaa Hammouda: Qabas: An Open-Source Arabic Lexicographic Database. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 13363–13370, Torino, Italia. ELRA and ICCL.

Mustafa Jarrar, Diyam Akra, Tymaa Hammouda: **ALMA: Fast Lemmatizer and POS Tagger for Arabic.** In Proceedings of the 2024 AI in Computational Linguistics (ACLING 2024), Procedia Computer Science, Dubai. ELSEVIER.

Tymaa Hammouda, Mustafa Jarrar, Mohammed Khalilia: **SinaTools: Open Source Toolkit for Arabic Natural Language Understanding.** In Proceedings of the 2024 AI in Computational Linguistics (ACLING 2024), Procedia Computer Science, Dubai. ELSEVIER.

Jarrar, M. (2021). **The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content.** Applied Ontology Journal, 16:1, 1-26. IOS Press.

Jarrar, M., & Amayreh, H. (2019). **An Arabic-Multilingual Database with a Lexicographic Search Engine**. In Proceedings – 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019). Lecture Notes in Computer Science (vol. 11608, pp. 234-246). Springer. Doi:10.1007/978-3-030-23281-8_19