



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

Unlike

**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،  
علمتين، علماء،  
العاليات  
وعلمائنا  
بعلمائكم

Scientist

Scientists

....

Inflectional Morphology will be the focus of this lecture

# Derivational Morphology (اشتقاق)

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

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

# (الاشتقاق بالإنجليزية) English Derivational Morphology

Examples of derivational patterns in English

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

....

...

....

# Arabic Derivational Morphology (الاشتقاق بالعربية)

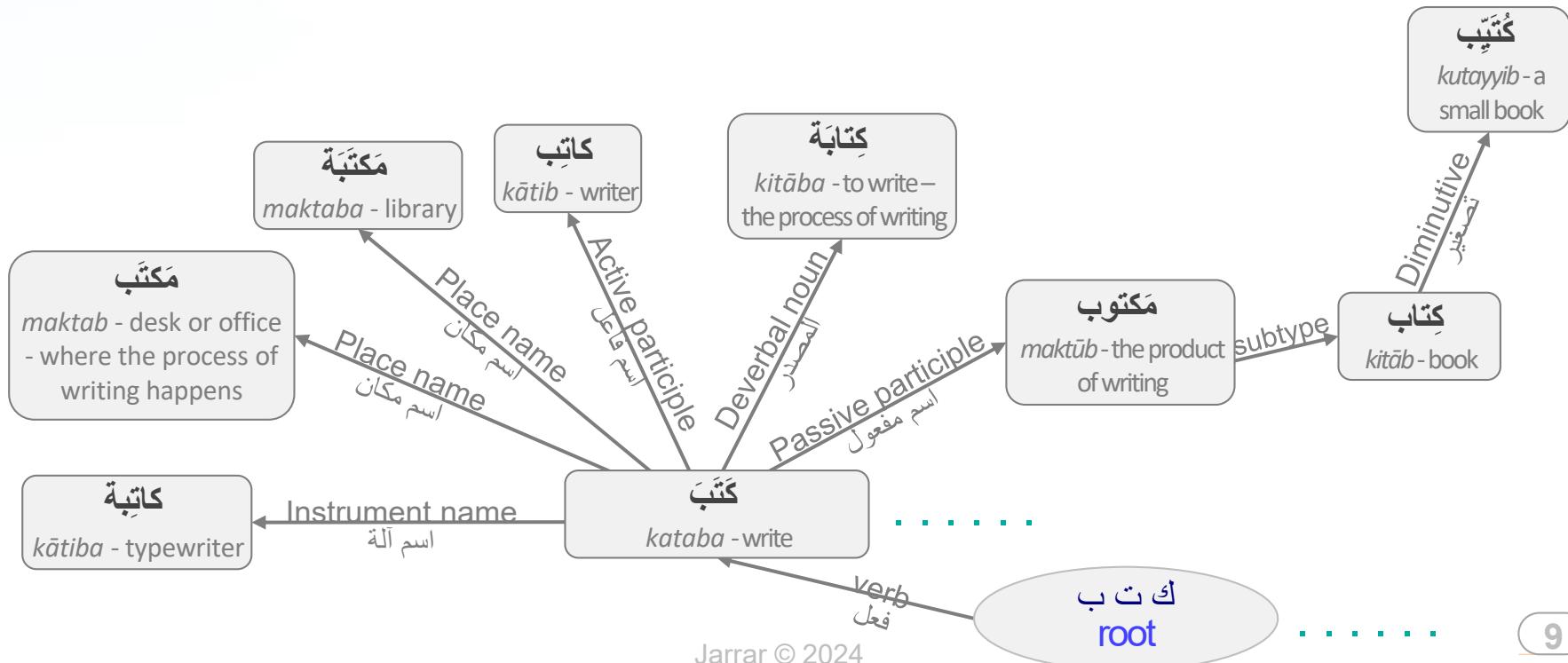
Derivational morphology in Arabic is

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

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

# الاشتقاق بالعربية (Arabic Derivational Morphology)

- Derivational morphology 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



# (لکسیم/وحدة صرفية) Lexeme

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

**Verb Lexemes:**

{write, writes, wrote, writing}

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

**Noun Lexemes:**

{scientist, scientists}

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

# (مدخلة معجمية) Lemma

**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**/اصطلاحاً to refer to a lexeme.

**Verb lemmas:** typically the past singular masculine 3rd form (الماضي المفرد المذكر الغائب)

{write, writes, wrote, writing}

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

**Noun Lemmas:** typically the singular form (المفرد)

{scientist, scientists}

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

# (مدخلة معجمية) Lemma

## Issues about lemmas

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

- Some lexicons split lemmas based on the common-use (شيوخ 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 (تفاحة، جامعة)
  - Plural or dual forms (ثقلان، اثنان، أختنان) or (تخيوم، تسعينيات، مُذهبات) 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
وسيكتها	كَتَبَ
إكتبلي	كَتَبَ
مكتباتهم	مَكْتَبَةٌ
مكاتبكم	مَكْتَبٌ
بمكتباتكن	مَكْتَبَةٌ
فكتابها	كِتَابٌ
كتابي	كِتَابٌ
كتيبنا	كُتْيَبٌ
فيكتاباتنا	كِتابَةٌ
بكتبة	كَاتِبٌ
وكتابهم	كَاتِبٌ
كتب	كِتابٌ

Word	Lemma
أبيات	بَيْت٢
بيوت	بَيْت٢ بَيْت١
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>



News Team Resources

ALMA (المى)

Arabic Morphology Tagger

Lemmatizer, POS tagger, and root tagger.

Accuracy: POS (93.8%), lemmatization (90.48%), and speed (32K tokens/second). Outperformed all other tools (see [article](#)).

أسرع المحللات الصرفية العربية وأكثرها دقة، وهو محلل مفتوح المصدر. يتم تحديد المدخلة المعجمية وقسم الكلام والجذر لكل كلمة في النص.

Morph Tagger

Lemmatizer

POS Tagger

- Downloads

Download [SinaTools](#) (Morph Module), you can also access ALMA memory (morphological solution ordered by frequency) which is part of SinaTools.

# Lemmatization

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

<https://sina.birzeit.edu/alma>

The screenshot shows the homepage of the ALMA Arabic Morphology Tagger. At the top left is the SinaLab logo, which consists of three colored circles (orange, green, blue) connected by lines. To the right of the logo is the word "SinaLab". On the far right of the header are three links: "News", "Team", and "Resources". Below the header, the word "ALMA" is written in large, bold, black letters, followed by "المى" in smaller black letters. Underneath "ALMA" is the text "Arabic Morphology Tagger". Below that is a short description: "Lemmatizer, POS tagger, and root tagger." Further down, it says "Accuracy: POS (93.8%), lemmatization (90.48%), and speed (32K tokens/second). Outperformed all other tools (see [article](#)).". A callout box contains the text: "أسرع المحللات الصرافية العربية وأكثرها دقة، وهو محلل مفتوح المصدر. يتم تحديد المدخلة المعجمية وقسم الكلام والجذر لكل كلمة في النص." At the bottom of the page are three buttons: "Morph Tagger", "Lemmatizer", and "POS Tagger".

- Downloads

Download [SinaTools](#) (Morph Module).

# Qabas

## Open-source Lexicographic Database

<https://sina.birzeit.edu/qabas>

The screenshot shows the homepage of the Qabas Lexicon. At the top right, there are language selection buttons for "Arabic" and "English". Below them is a "About" link. The central feature is the Qabas logo, which consists of three colored circles (orange, teal, and light blue) connected to a white circle containing the Arabic letter "ق". To the right of the logo, the text "معجم قبس الحاسوبى" and "Qabas Lexicon" is displayed. Below the logo is a search bar with the placeholder text "Search for a term ...". Underneath the search bar, the text "Most Commonly Used Words: [كلمة](#), [معجم](#), [قبس](#)" is shown. A link "Browse: [Ontology Tree](#)" is also present. At the bottom of the page, there is a footer with the Birzeit University logo and the text "BIRZEIT UNIVERSITY Copyright © 2024 Birzeit University Jatta © 2024".

Arabic | English

About

Search for a term ...

Most Commonly Used Words: [كلمة](#), [معجم](#), [قبس](#)

Browse: [Ontology Tree](#)

BIRZEIT UNIVERSITY  
Copyright © 2024 Birzeit University  
Jatta © 2024

# Qabas

## Open-source Lexicographic Database

The screenshot shows the Qabas lexicographic database interface. At the top, there is a search bar with the text 'كلمات' (Words) and a magnifying glass icon. Below the search bar are buttons for 'Translations', 'Synonyms', and 'Definitions'. The main content area displays the search results for the word 'كلمة' (Word). The results include:

- اللغة: فصحي حديثة
- العدد: مفرد
- الجذر: ك ل م
- قسم الكلام: اسم
- الجنس: مؤنث

Below the search results, there are two tabs: 'معاني و مدخلات في المعاجم العربية' (Meanings and entries in Arabic dictionaries) and 'سياقات في مدونات نصية' (Contexts in textual annotations). The 'معاني و مدخلات...' tab is currently selected. It lists several definitions and examples:

- + كلمة | **كلمة** (معجم المعاصرة)
- + كلمة 1 (معجم سما)
- كلمة (المعجم الفلسفية)
- في اللاهوت المسيحي ، الكلمة Verba هي الشخص الثاني في الثالوث المسيحي : الأب والابن والروح القدس.
- عند مالبراشن : كلمة الله هي المبدأ الإلهي والعقل الكلي الذي تنطوي في جوهرة الحقائق كلها .
- عند المتضوفة: ما يكتنف به عن الماهيات والحقائق الأزلية الثابتة ، وكلمة 'الحضره' هي صورة الإرادة الكلية ، وتشير إلى 'كن'، في قوله تعالى "إِنَّمَا أَمْرُهُ إِذَا أَرَادَ شَيْئاً أَنْ يَقُولَ لَهُ كَنْ فَيَكُونُ'.
- كلمة (الانتropolوچيا العربية)
- تعبير كتابي للإشارة أو الدلالة على معنى
- كلمة (الانتropolوچيا العربية)

On the right side of the interface, there is a sidebar titled 'About' which contains a list of related terms:

- كلام
- كلم
- كلام
- كلام
- كلام
- كلام
- كلام
- كلام | كرم
- كلاما
- كلاماني
- كلمة | **كلمة**
- كلامنة

<https://sina.birzeit.edu/qabas>

# Qabas

# Open-source Lexicographic Database

كلمات

Translations   Synonyms   Definitions

Ontology Dictionaries Qabas lexicon About

(0.01 secs)

# كلمة | كِلْمَة

الجذر: ك ل   قسم الكلام: اسم   اللغة: فصحي حديثة

الجنس: مؤنث   العدد: مفرد

بيانات في المعاجم العربية

مدونات فصحي: القرآن الكريم   مدونة سلمي   مدونة ATB-LDC

مدونات عامة: فلسطينية   ليبانية   سورية   مصرية   عراقية   لبيبة   سودانية   يمنية   إمارانية

كلمة وردت في 1201 سياق   80   528   151   84   78   40   33   29   26   24

+   +   +   +   +   +   +   +   -

اسم مفرد مؤنث   اسم مفرد

الكلمة التي فضحت نصف الشعب !

واحد بنوره ولا برد عليه لأن هذا نكروه وانسان جايف معنى الكلمة

كلمة | كِلْمَة

كلمة

كلمات

الكلمة

بكلمة

كلمات

الكلمات

كلمات

كلمات

كلمات

كلمات

كلمات

# Qabas

## Open-source Lexicographic Database

Search for a term ... Search

Translations   Synonyms   Definitions

Ontology   Dictionaries   Qabas lexicon

About

### معجم قبس الحاسوبي

معجم قبس يتكون من 58,000 مدخلة و تم ربطه ب 110 معاجم و 10 مدونات نصية

ربط 10 مدونات نصية

تم ربط مدخلات 10 مدونات نصية بمدخلات قبس

مدخلات مقابلة في 110 معاجم

تم ربط مدخلات 110 معاجم بمدخلات قبس

أخبار

- عن المعجم
- أخبار
- إحصائيات
- مقدمة المعجم
- شمولية معجم قبس
- mirratat\_tatbirat\_mugam
- فلسفة المعجم
- المصادر اللغوية
- الدليل المعياري
- تصنيف المدخلات
- أسئلة شائعة
- حقوق الملكية
- تنزيل قبس

MoU with ALECSO to digitize 50 dictionaries

Digitize dictionaries of the Academy of the Arabic language-Cairo

Shoman Arab Researchers Award

Bin Rashid's Arabic Language Award

More News

إحصائيات

إحصائيات للمدونات التي تم ربطها مع قبس

إحصائيات للمعاجم التي تم ربطها مع قبس

إحصائيات حسب قسم الكلام

The screenshot shows the homepage of the Qabas website. At the top, there is a search bar with a magnifying glass icon and a small globe icon. Below the search bar are three filter buttons: 'Translations', 'Synonyms', and 'Definitions'. The main navigation menu includes 'Ontology', 'Dictionaries', 'Qabas lexicon', and 'About'. The central part of the page features a large title 'معجم قبس الحاسوبي' (Qabas Lexicographic Database) with a subtitle indicating it contains 58,000 entries and is linked to 110 other dictionaries and 10 digital encyclopedias. Below this, there are two sections: one titled 'ربط 10 مدونات نصية' (Linking 10 digital encyclopedias) and another titled 'مدخلات مقابلة في 110 معاجم' (Interviews in 110 dictionaries). To the right, a sidebar lists various statistical and informational links. At the bottom, there are four thumbnail images representing different partnerships or awards, with labels below them: 'MoU with ALECSO to digitize 50 dictionaries', 'Digitize dictionaries of the Academy of the Arabic language-Cairo', 'Shoman Arab Researchers Award', and 'Bin Rashid's Arabic Language Award'. A 'More News' button is located at the bottom left of this section. The footer contains a logo for 'BIRZEIT UNIVERSITY' and the text 'Copyright © 2024 Birzeit University'.

# 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, (سن و) (سن ي) سنة

**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	كتبـهـا	ك ت بـ	كتبـهـا
مكتباتهم	مـكـتبـةـهـمـ	ك ت بـ	مـكـتبـهـمـ
مـكـاتـبـكـمـ	مـكـتبـكـمـ	ك ت بـ	مـكـاتـبـكـمـ
كتابـيـ	كتـابـيـ	ك ت بـ	كتـابـيـ

# (زوائد صرفية) Inflectional Affixes

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

يكتب، سـيكتب، وـسيكتب، الـكتاب، فـكتب، أـكتب، كـالكتاب،

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

book**s**, book**ed**, happ**iness**, usually, help**ful**, ...  
كتب**ت**، كتب**ها**، كتب**لك**، كتب**كم**، كتب**هن**، ....

**Infixes (أواسط):** inserted inside the stem to form a word form.

اجتهـد، ابـتـاع، ...

**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، أكتب، فكتب، الكتاب، 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:

# Suffixes (لواحق)

attached before the stem to form a word form.

**books**, **booked**, **happiness**, **usually**, **helpful**, ...

كتب، كتبها، كتبلك، كتبكم، كتبكما، كتبهن، ...

# In English:

- changes the grammatical properties of a word within its syntactic category
  - 8 suffix types: Verbs{makes, booked, eating, eaten}, Nouns{girls, oxen}, adjectives and Adverbs {larger, largest}.
  - Irregulars: sheep, children, ran, drunk, ....

## In Arabic:

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

# 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 syntactic 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