



On Arabic Stop-words: A Comprehensive List and a Dedicated Morphological Analyzer

Driss Namly, Karim Bouzoubaa, Rachida Tajmout, Ali Laadimi

ALELM Team



Stop-words

"كاسبرسكي" ترصد 5 ملايين هجمة إلكترونية بالمغرب في 3 أشهر

سيبريس

أفادت معطيات صادرة عن شركة "كاسبرسكي"، المتخصصة في مكافحة البرامج الضارة وجرائم الإنترنت، بأن عدد الحوادث المتعلقة بالبرامج الخبيثة والضارة التي تم تنزيلها من الأنترنت في المغرب، ما بين أبريل ويونيو 2019، فاقت 5 ملايين.

وجاء في نشرة إحصائية للشركة حول الأمن السيبراني بالمغرب أن حوالي 30.7 في المائة من مستخدمي برنامج "كاسبرسكي" للأمن كانوا ضحية هذه التهديدات عبر الإنترت، وهو ما يجعل المغرب في المرتبة 34 عالمياً من حيث الدول المتأثرة بالتهديدات السيبرانية المتعلقة بتصفح "الويب".

وسُجل في الجزائر أيضاً خلال الفترة سالفة الذكر تعرض المستخدمين لمثل هذه التهديدات عبر الإنترت بنسبة تصل إلى 44 في المائة، وبنسبة 71.3 في المائة في أفغانستان، وبنسبة 69 في المائة في طاجيكستان.

وقال باسكال نودين، وهو مسؤول في شركة "كاسبرسكي" في شمال إفريقيا، إن الاستنتاجات الخاصة بهذه الفترة تدفع إلى "الوصيصة مرة أخرى للعمل على زيادة الوعي بين المغاربة لاطلاعهم على الدور الرئيسي لأمن تكنولوجيا المعلومات"، متربماً إلى أن "غالبية التهديدات المسجلة لدى المغاربة يمكن معالجتها من خلال إجراءات بسيطة يتوجب تملكها من لذهم".

وخلال الربع الثاني من السنة الجارية، نجحت برنامج "كاسبرسكي" للأمن في كشف أكثر من 15 ألف برنامج خبيث في أجهزة الكمبيوتر التي تشغّل "Kaspersky Security Network"، ما يضع المغرب في المرتبة 61 على المستوى العالمي في هذا الصدد.

وفي نظر خبراء "كاسبرسكي"، فإن الحماية من مثل هذه الأخطار لا تتطلب فقط تثبيت برامج مكافحة الفيروسات القادر على معالجة الملفات المصابة، بل أيضاً التوفير على جدار حماية، إضافة إلى ميزة مكافحة الجذور الخبيثة (anti-rootkits)، وهي عملية خفية تغير سلوك نظام التشغيل أو نوائه ويسمح لها بالتوارد في نظام الضحية لشهر أو أحياناً سنوات.

% 32.02 = 253 / 81

Definition

- Common words that frequently appear
- Isolated
 - no information
- In a sentence
 - grammatical and syntactical function

Stop-words exploitation & removal

Google search results for "دخل الولد الى القسم":

- Washington Open Adoption - Journeysoftheheart.net**
Annonce www.journeysoftheheart.net/ ▾
Domestic Program For Birthmothers. Place baby or toddler at our agency
- موسوعة اللغة العربية - نماذج في الإعراب**
www1.amalnet.k12.il/...%20في%20الإعراب.aspx ▾ Traduire cette page
نماذج في الإعراب على آخره. دخل إلى الصفت. دخل الولد - 17 oct. 2012 : فعل ماض مبني على الفتح الظاهر على آخره. دخل إلى الصفت. دخل الولد : فاعل مرفوع وعلامة رفعه الضمة الظاهرة على آخره . إلى : حرف جر مبني على ...
- نكت مغربية - قالك هادا واحد الولد دخل القسم معطل. وقاله...**
www.facebook.com/permalink.php?id...story... ▾ Traduire cette page
قالك هادا واحد الولد دخل القسم معطل. وقاله الأستاذ مالك تعطلت قال الوليد كان كايضرب الوليدة . قاله الأستاذ مالك ماجيني وخليه يضربيها. قاله الولد راه...
- اعرب هذه الجمله ذهب الولد الى المدرسه ؟؟ - نابليسيه ولاد جبل ...**
<https://www.facebook.com/Jabal.../33209945019556...> ▾ Traduire cette page
Likes23 26 اعرب هذه الجمله ذهب الولد الى المدرسه ؟؟ ... ذهب الولد الى المدرسه Comments · Like Comment. Sabah Saleh, Salmane Mahmoude Zakariya, Zyad ... Mohamed
- نماذج في الإعراب - منتديات الدراسة الجزائرية**
www.eddirasa.com/forum/t551/ ▾ Traduire cette page
28 déc. 2013 - 9 messages - 6 auteurs منتديات الدراسة الجزائرية > منتديات التعليم الإبتدائي > قسم الرابعة و الخامسة ابتدائي ... دخل: فعل ماض مبني على الفتح. ... دخل الولد إلى الصفت.

IR +

- Text categorization
- Information extraction
- Text summarization
- Search engine indexing
- ...

Removal

- corpus size --
- Saves space in indexes
- Increases accuracy without damaging retrieval efficiency

Stop-words identification

Static stop-words list

- Corpus-dependent
- Rely on the frequency feature
- Don't consider all clitized forms

contains هو

excludes clitized forms ... وهو، فهو، ...



Analysis

To remove these words

we must identify them

List



Morphological analysis

- The grammatical category default
- Don't consider all clitized forms

detect هو (POS=particle)
ignore الذي (POS=noun)

Objectives

01



A comprehensive Arabic rule-based stop-words list

02



A contextual stop-words analyzer

State of the art

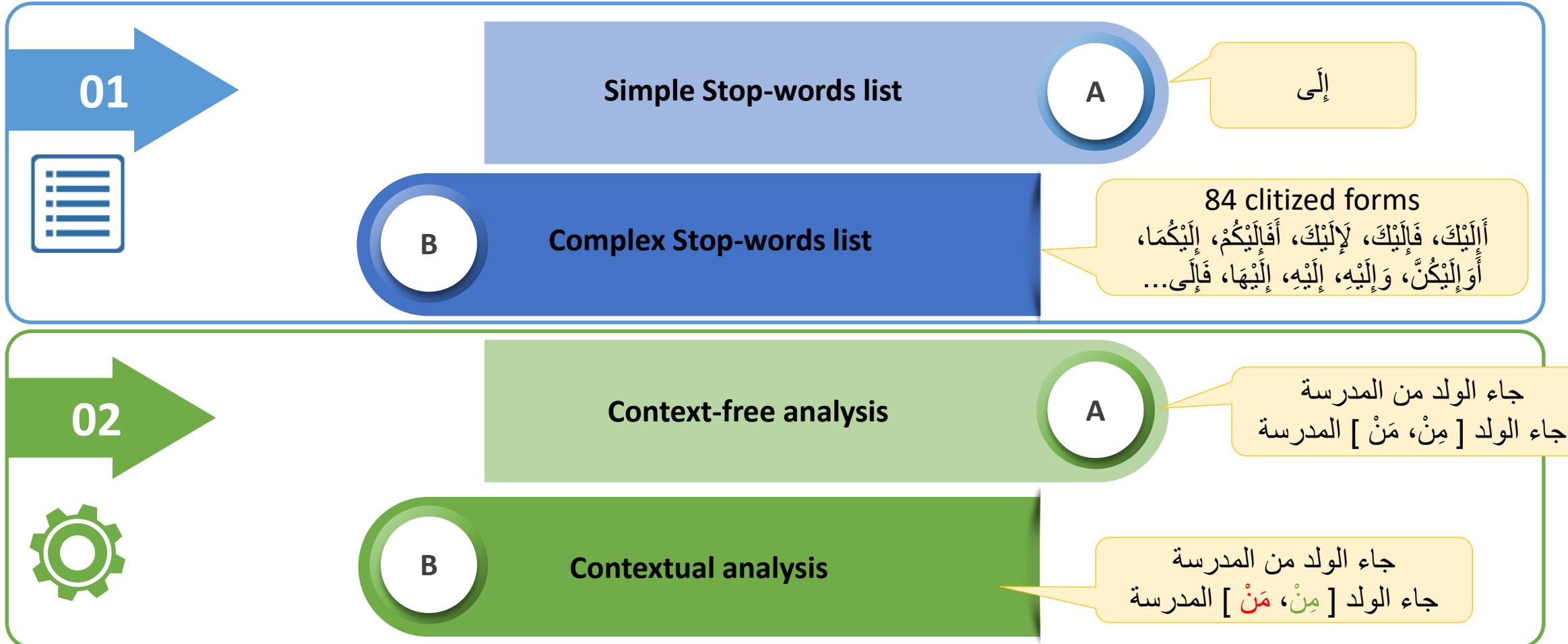
Stop-word lists

- Shereen Khoja (168)
- Ibrahim Abu El-Khair(1529)
- Walaa Medhat et al. (1061)
- Alajmi et al. (200)
- Stop words project(162)
- Ranks NL(102)

Morphological analysis

- BAMA
- ALKHALIL
- MADAMIRA
- CALIMA-star

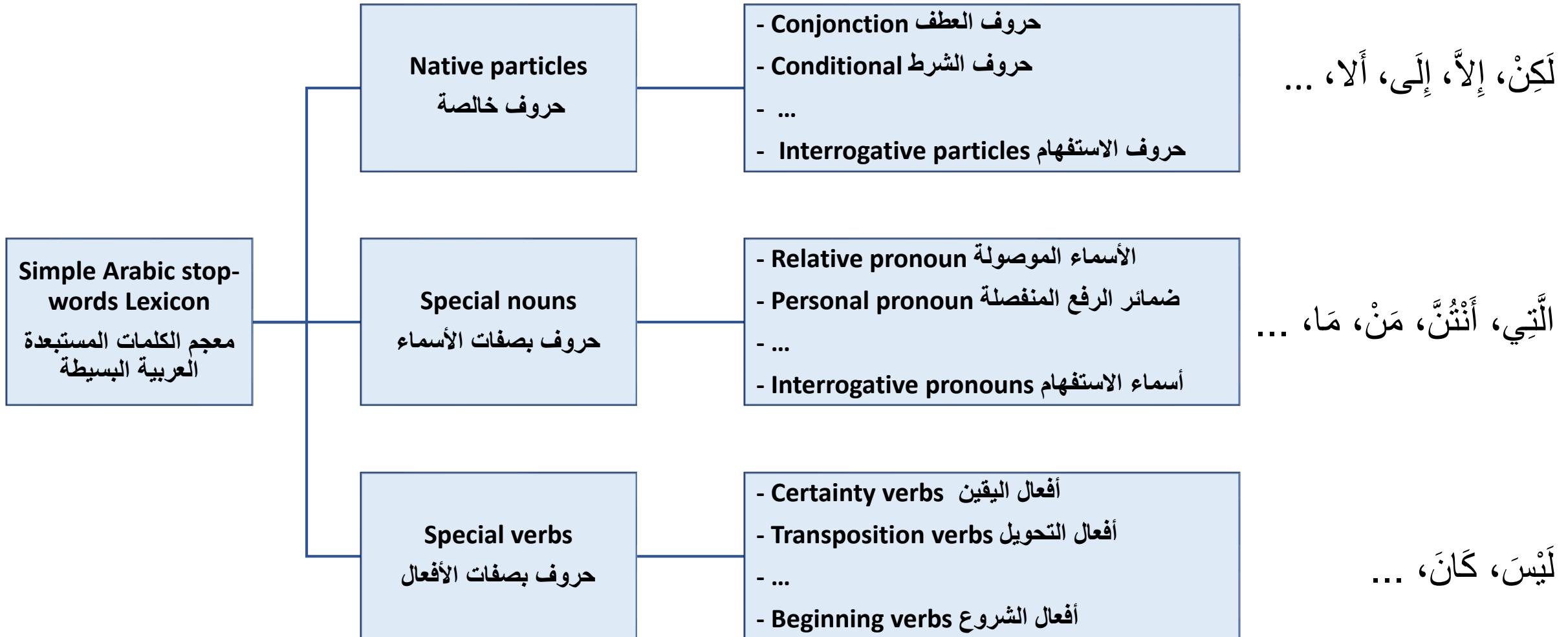
Proposed approach



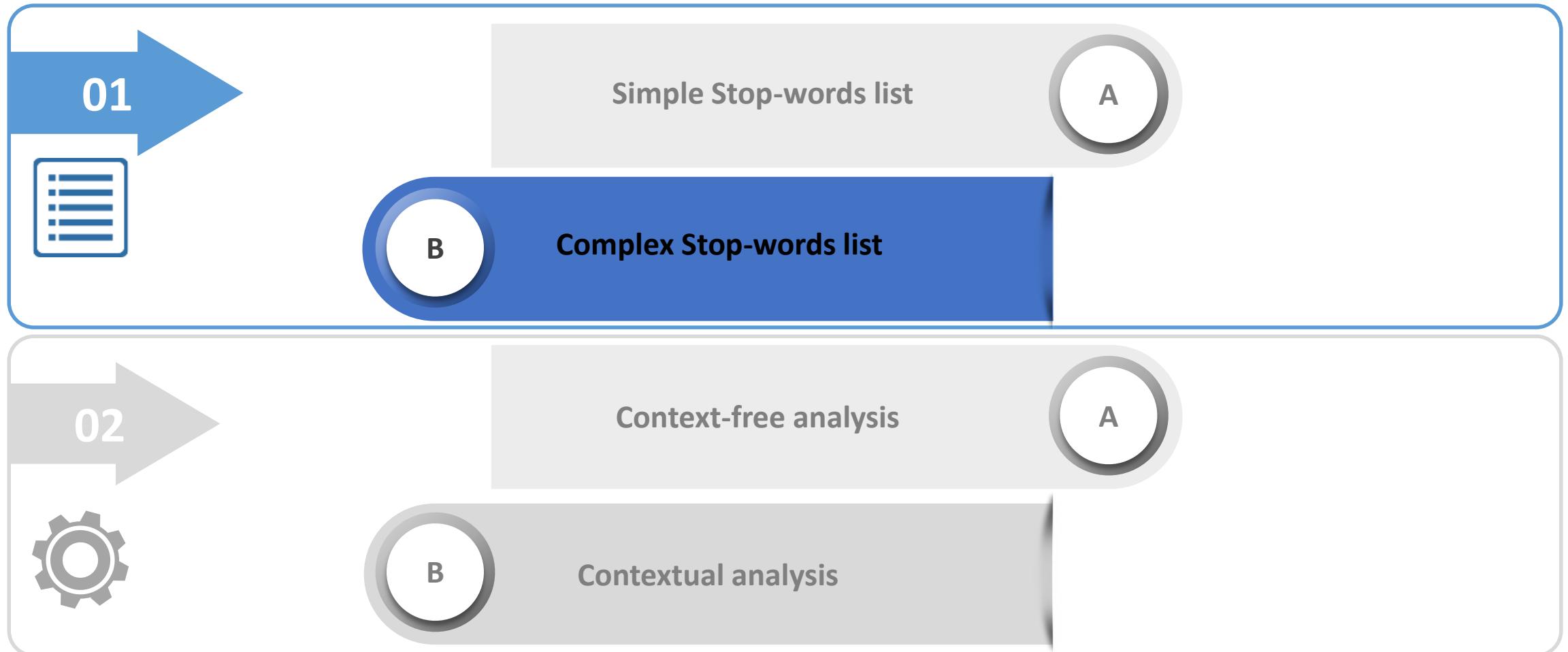
Proposed approach



1.A- Simple Stop-words list design

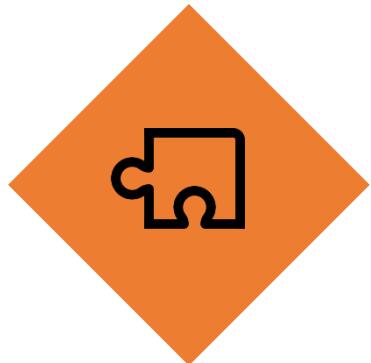


Proposed approach



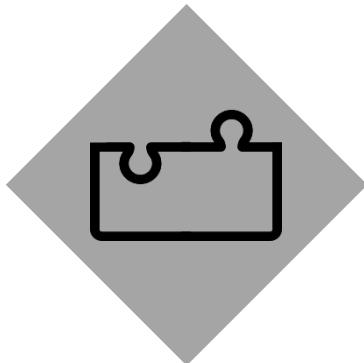
1.B- Complex Stop-words list design

ProClitic



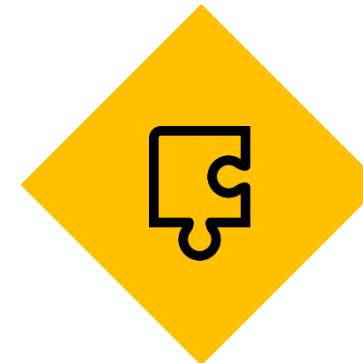
فَبِـ

Simple Stop-Word



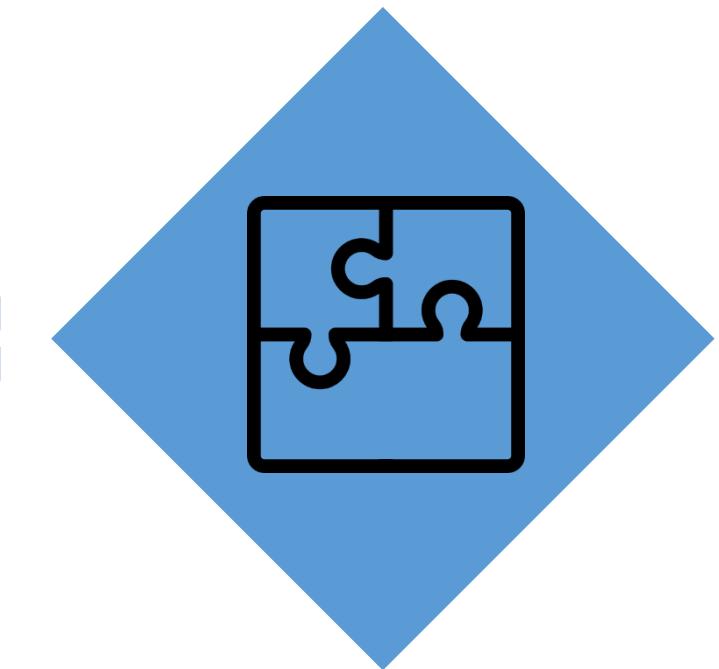
غَيْرـ

EnClitic



هُمـ

Complex Stop-Word



فِـغَيْرِـهِمـ

1.B- Complex Stop-words list design

ProClitic

Conjunction	حروف العطف
Prepositions	حروف الجر
Accusative Particles	حروف النصب
Future particle	سين المستقبل
Definition particle	التعريف
...	...

EnClitic

connected pronouns	الضمائر المتصلة
prevention	نون الوقاية
Ascribing	ياء النسبة
...	...

Arabic rules

ال + حرف

Word type

Particle	حرف
Noun	اسم
Verb	فعل

Arabic rules

اسم موصول + ضم

Proposed approach



1- Arabic Stop-words List (ASL)

diacritized stop-words list named ASL (Arabic Stop-words List) :

Enclitic	Simple stop-word		Proclitic	Complex stop-word		Category
-	maybe	عَلَّ	وَلَ	And maybe	وَلَعَلَّ	NP
-	that	ذَلِكَ	كَ	like that	كَذَلِكَ	SN
-	you are	تَكُونِينَ	سَ	we will be	سَتَكُونِينَ	SV
كِ	no more than	غَيْرُ	-	no more than you	غَيْرُكِ	SN
كِ	maybe	عَسَى	-	Maybe you	عَسَاكِ	SV
هُمْ	to	إِلَى	أَوْ	And is to they	أَوْ إِلَيْهِمْ	NP

	Native Particles	Special nouns	Special verbs	Total
Simple list	84	263	1,784	2,131
Complex list	1,590	9,627	27,936	39,153

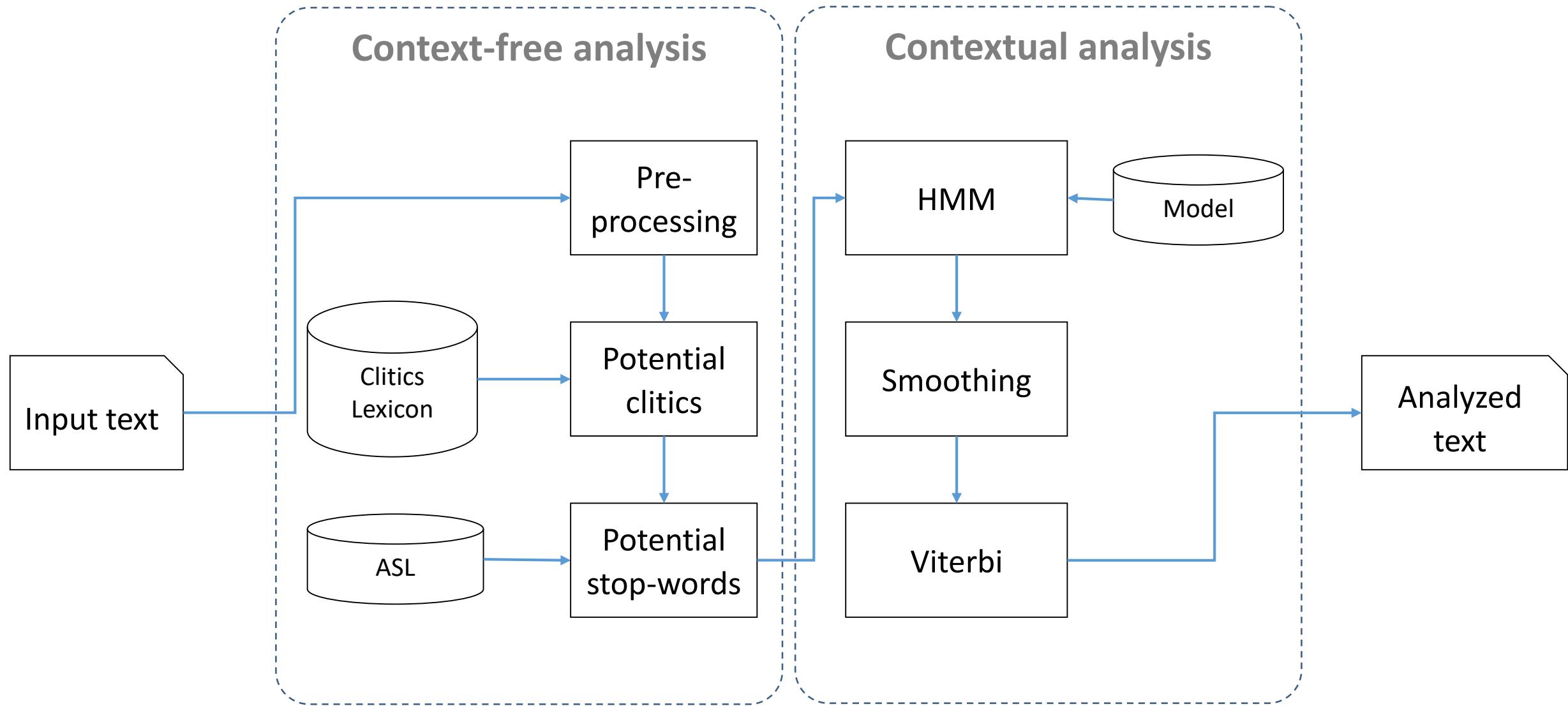
كُنْتَ، كُنَّا، كُنْتُمَا، كَانَتْ،
كَانُوا، نَكُونُ، ...

Khoja (168)
Abu El-Khair(1529)
Alajmi (200)

Proposed approach

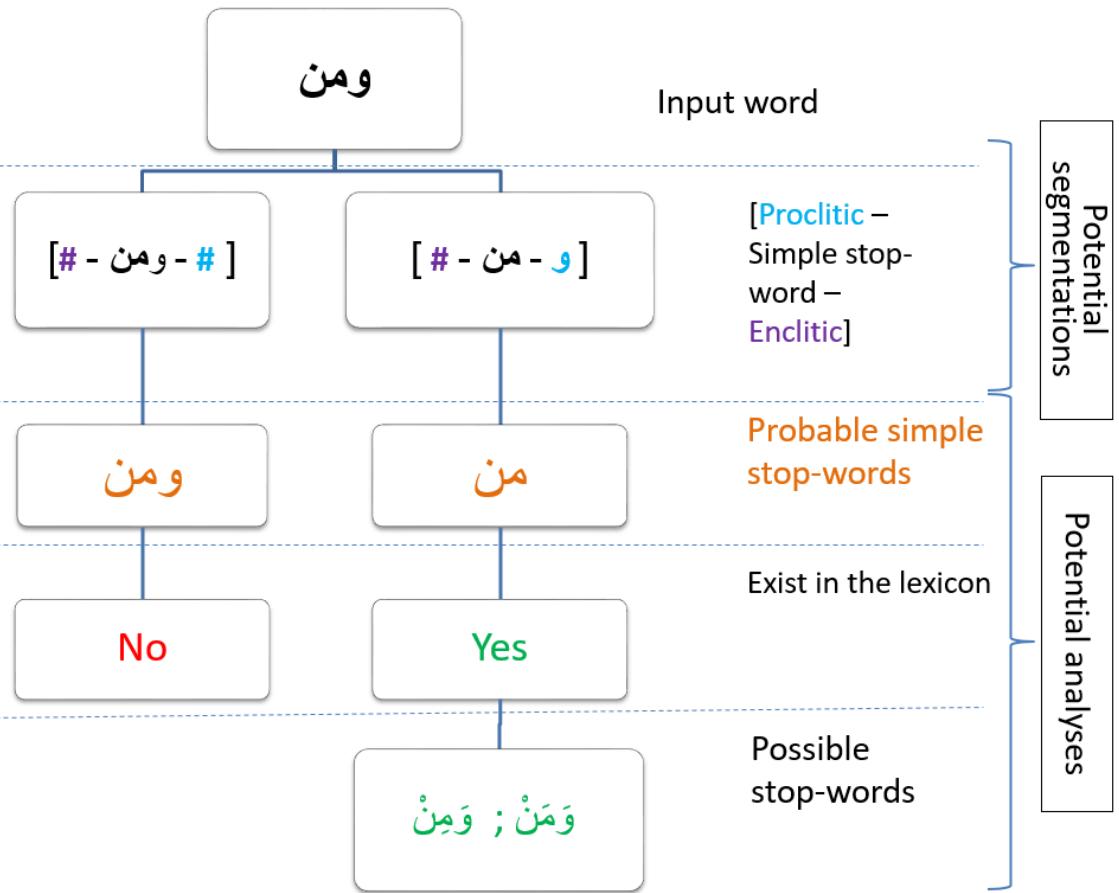


2- Stop-words analyzer design

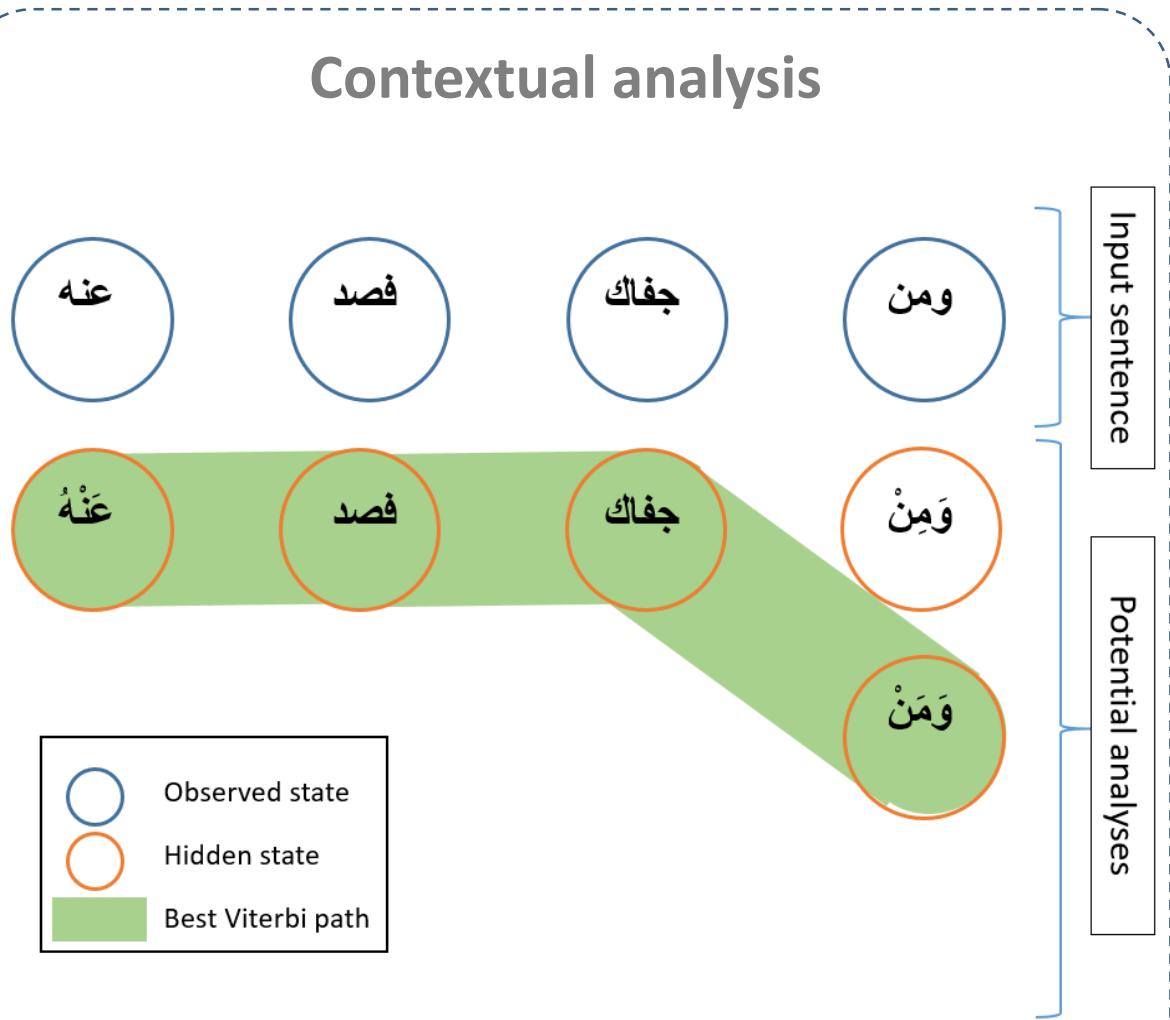


2- Stop-words analyzer design

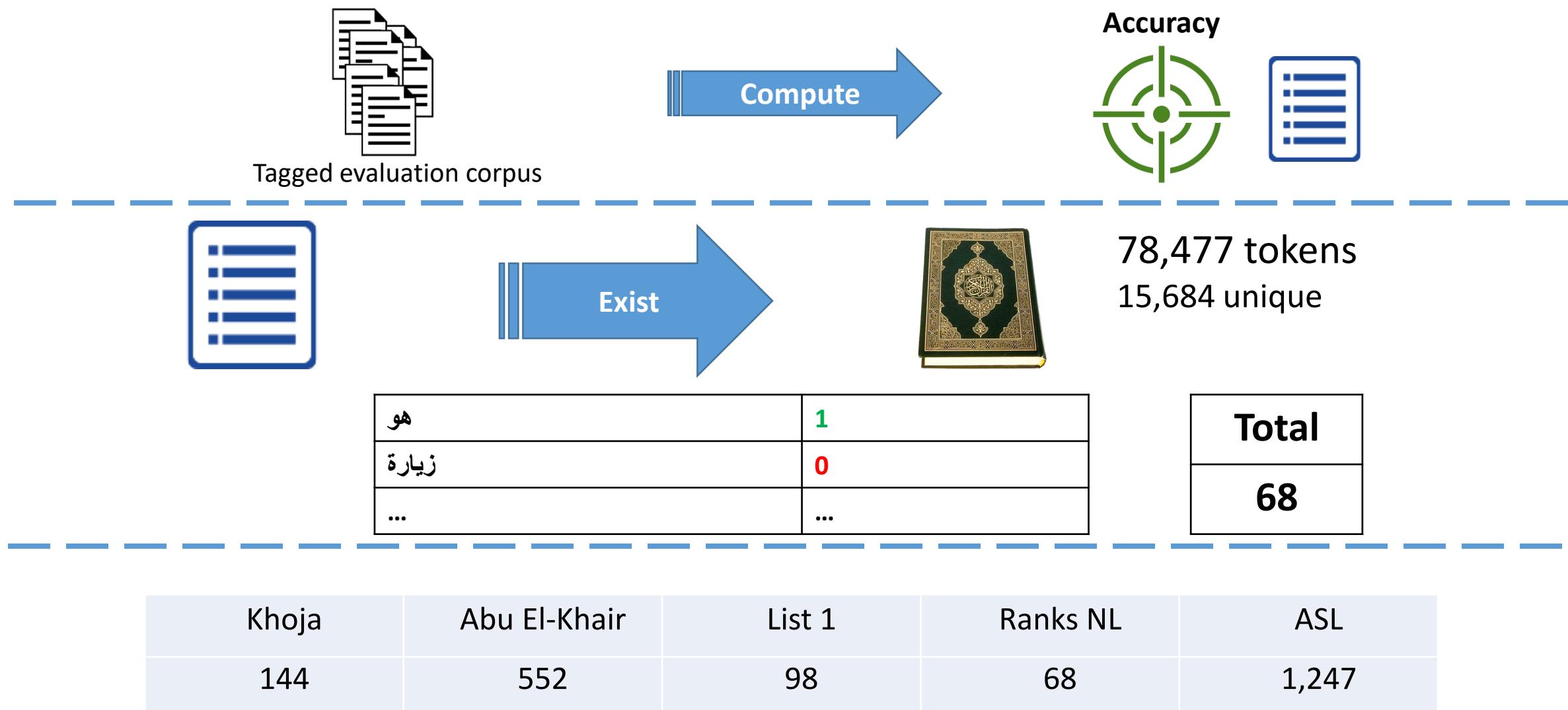
Context-free analysis



Contextual analysis



Stop-words list evaluation



Context-free evaluation



Corpus
(manually annotated)
1,628 words
419 stop-words



Accuracy



تتفرد المدينة العتيقة بموهّلات متنوعة ولا
محدودة، انضهرت فيها روافد ثقافية متباينة
ميزتها عن غيرها من المدن

تنفرد المدينة العتيقة بمؤهلات متنوعة [وَلَا] محدودة،
انصهرت [فيها] روافد ثقافية متباينة ميزتها [عَنْ، عَنْ، عَنْ]
[غَيْرِهَا، غَيْرِهَا] [مَنْ، مُنْ، مِنْ، مَنْ] المدن

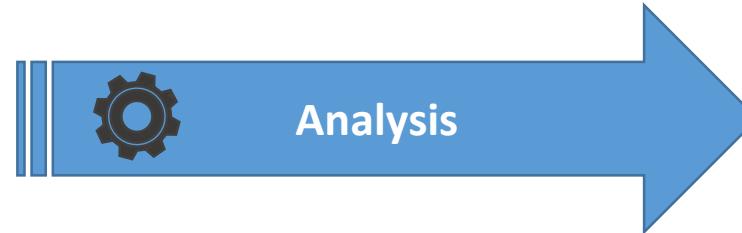
94.11%

BAMA	Alkhalil 1	Alkhalil 2	CALIMA-star	ASL analyzer
91.09	88.70	93.98	93.12	99.94

Contextual evaluation



Corpus
(manually annotated)
1,628 words
419 stop-words



تنفرد المدينة العتيقة بمؤهلات متنوعة ولا
محدودة، اانصهرت فيها روافد ثقافية متباينة
ميزتها عن غيرها من المدن

تنفرد المدينة العتيقة بمؤهلات متنوعة [وَلَا] محدودة،
انصهرت [فِيهَا] روافد ثقافية متباينة ميزتها [عَنْ، عَنْ، عَنْ]
[غَيْرَهَا، غَيْرَهَا] [مَنْ، مُنْ، مِنْ، مَنْ] المدن

82.35%

MADAMIRA

90.17

ASL analyzer

97.85

Conclusion

Identification

- Static list
- Morphological analysis

Default

- grammatical category
- clitized forms

Designed

- comprehensive Arabic rule-based stop-words list
- contextual stop-words analyzer

Evaluation

- 1,247 Vs 552 (2nd list)
- context-free 99.94%
- contextual 97.85%



Thank you

On Arabic Stop-words: A Comprehensive List and a
Dedicated Morphological Analyzer

Driss Namly, Karim Bouzoubaa, Rachida Tajmout, Ali Laadimi

ALELM Team

