Université Mohammed V de Rabat

ECOLE MOHAMMADIA D'INGENIEURS

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

Driss Namly, Karim Bouzoubaa, Rachida Tajmout, Ali Laadimi

ALELM Team

ICALP 2019

16-17 Octobre- Nancy

Stop-words

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

هسبريس

أفادت مُعطيات صادرة عن شركة "كاسبرسكي"، المتخصصة في مكافحة البرامج الضارة وجرائم الإنترنت، بأن عدد الحوادث المتعلقة بالبرامج الخبيثة والضارة التي تم تنزيلها من الأنترنيت 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 الإنترنت، 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 الإنترنت بنسبة تصل إلى 44 في المائة، وبنسبة 71.3 في المائة في أفغانستان، وبنسبة 69 في المائة في طاجكستان.

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

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

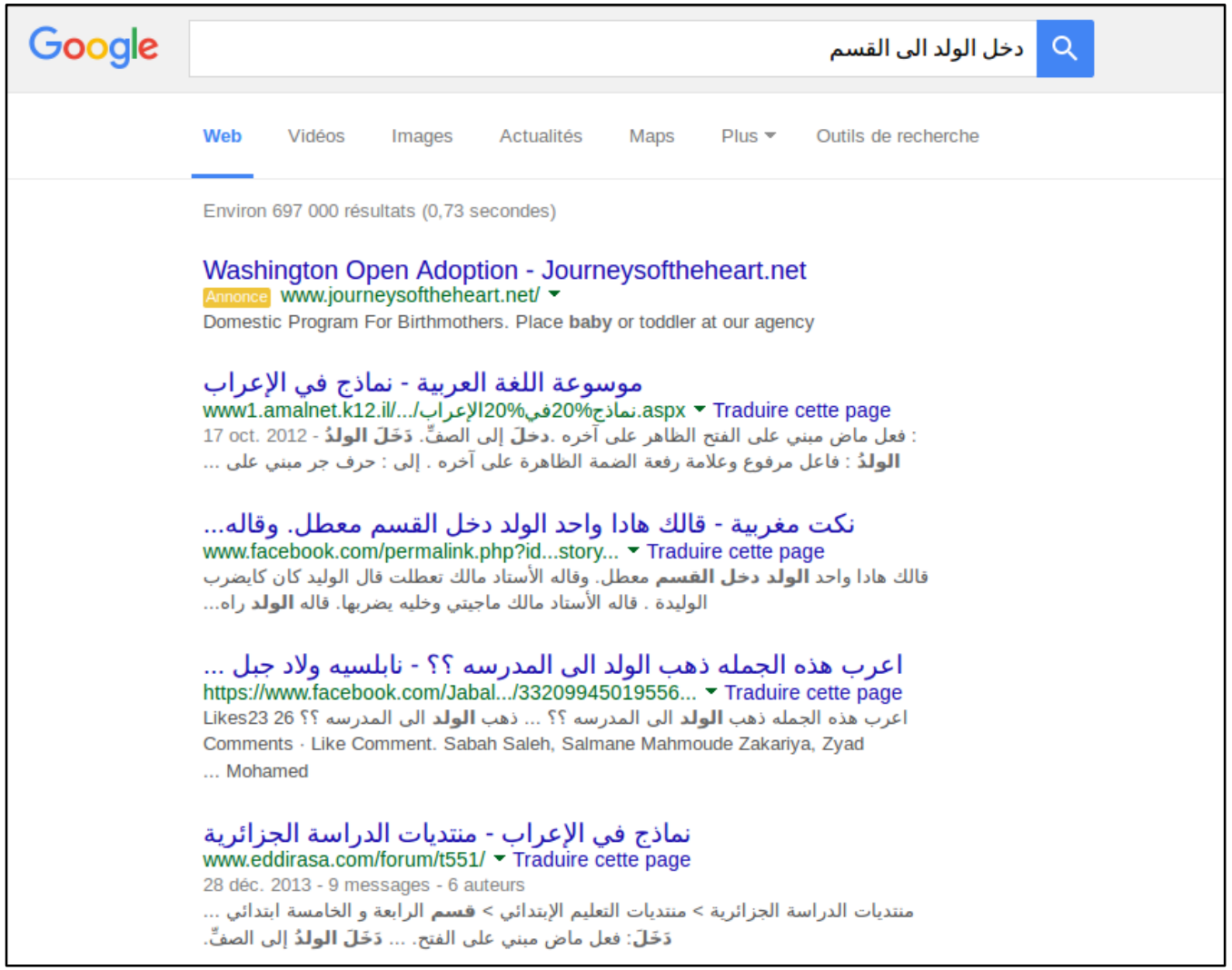
وفي نظر خُبراء "كاسبرسكي"، فإن الحماية من مثل هذه الأخطار لا تتطلب فقط تثبيت برامج مكافحة الفيروسات القادرة على معالجة الملفات المصابة، بل أيضاً التوفر على جدار حماية، إضافة إلى ميزة مكافحة الجذور الخفية (anti-rootkits)، وهي عملية خفية تغير سلوك نظام التشغيل أو نواته ويسمح لها بالتواجد في نظام الضحية لشهور وأحياناً لسنوات.

% 32.02 = 253 / 81

Definition

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

Stop-words exploitation & removal



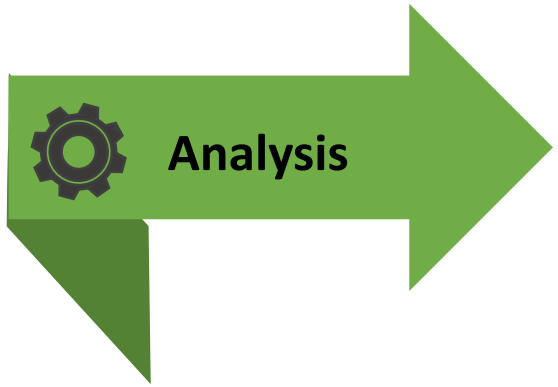
IR +

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

Removal

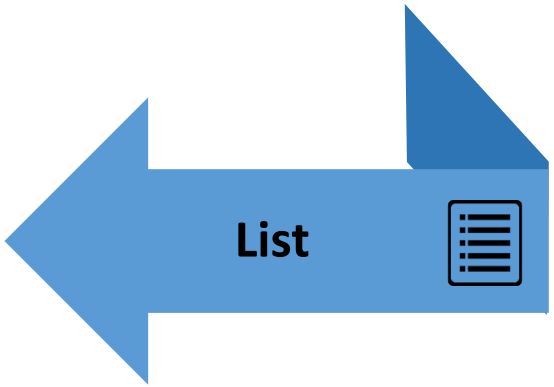
- corpus size --
- Saves space in indexes
- Increases accuracy without damaging retrieval efficiency

Stop-words identification



To remove these words

we must identify them



Static stop-words list

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

contains هو
excludes clitized forms ...وهو، فهو

Morphological analysis

- The grammatical category default
- Don’t c… …er all clitized for…

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

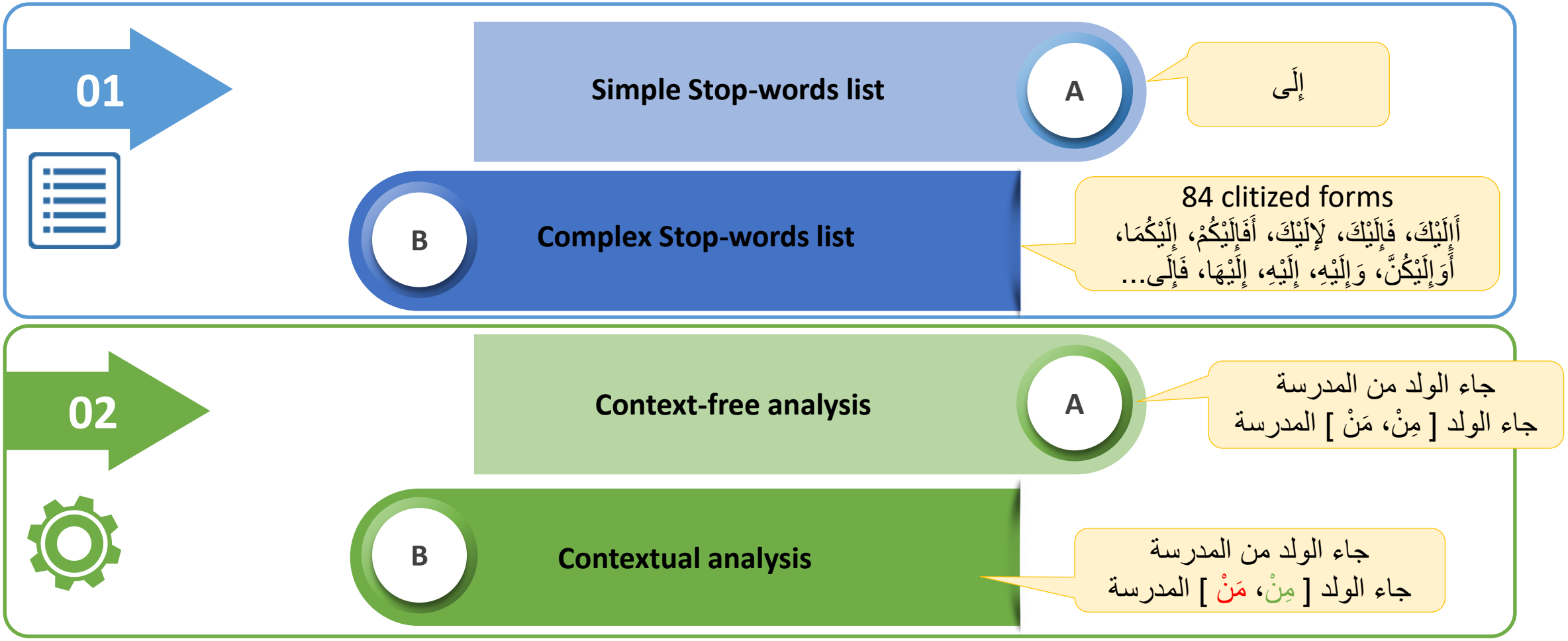
Stop-word lists

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

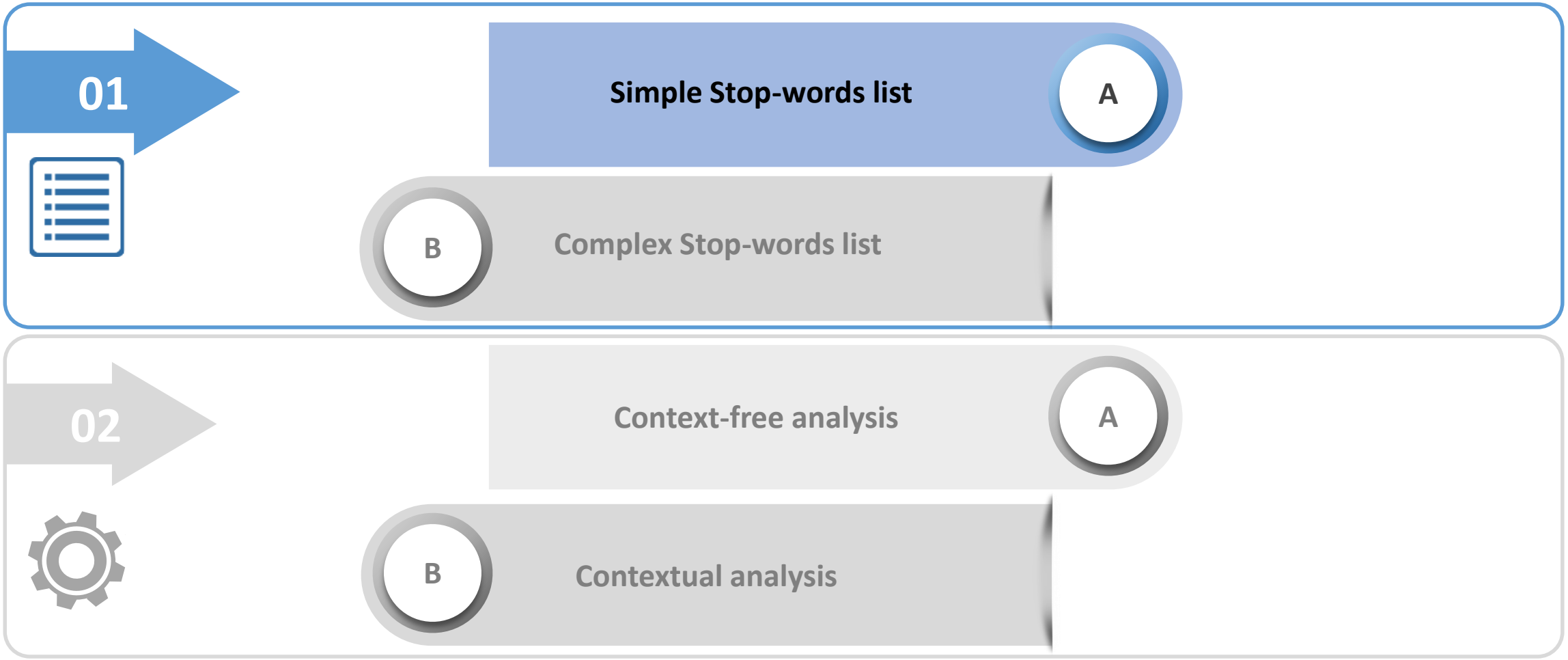
Morphological analysis

- BAMA
- ALKHALIL
- MADAMIRA
- CALIMA-star

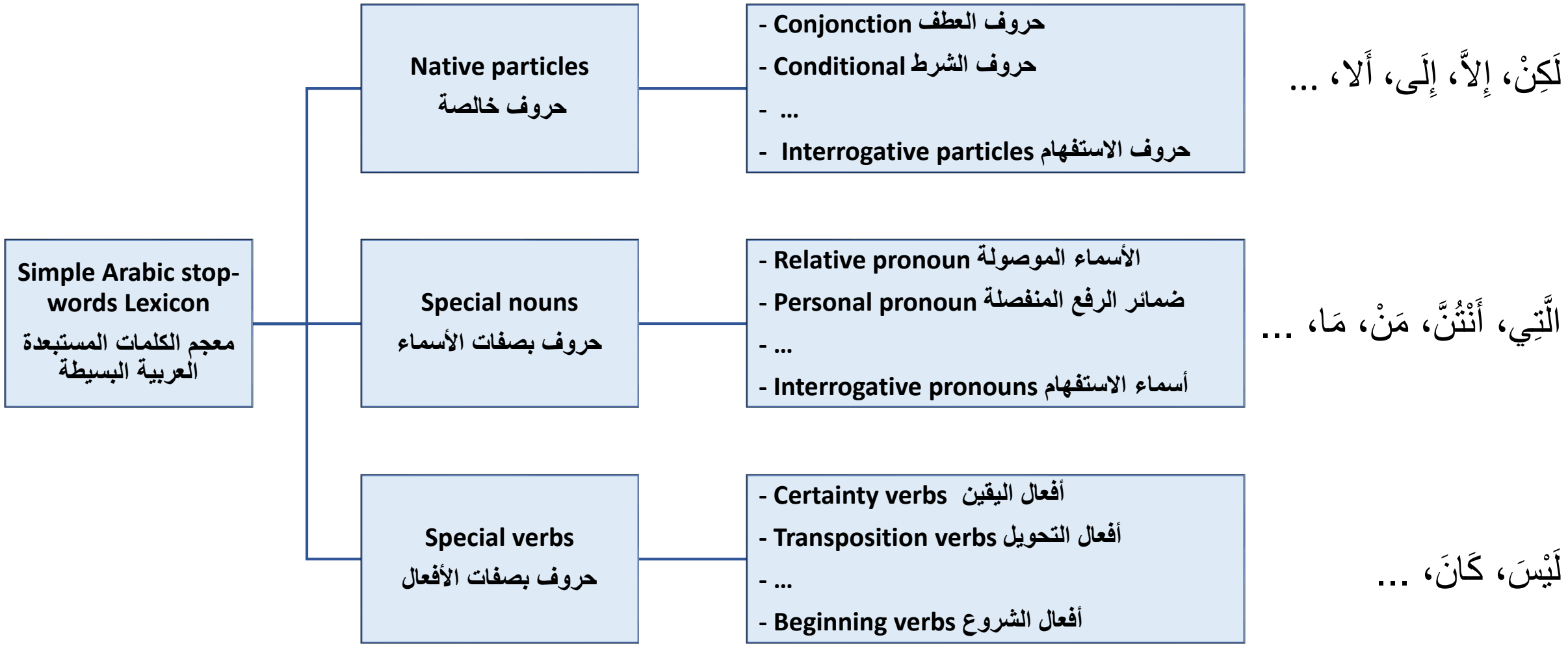
Proposed approach



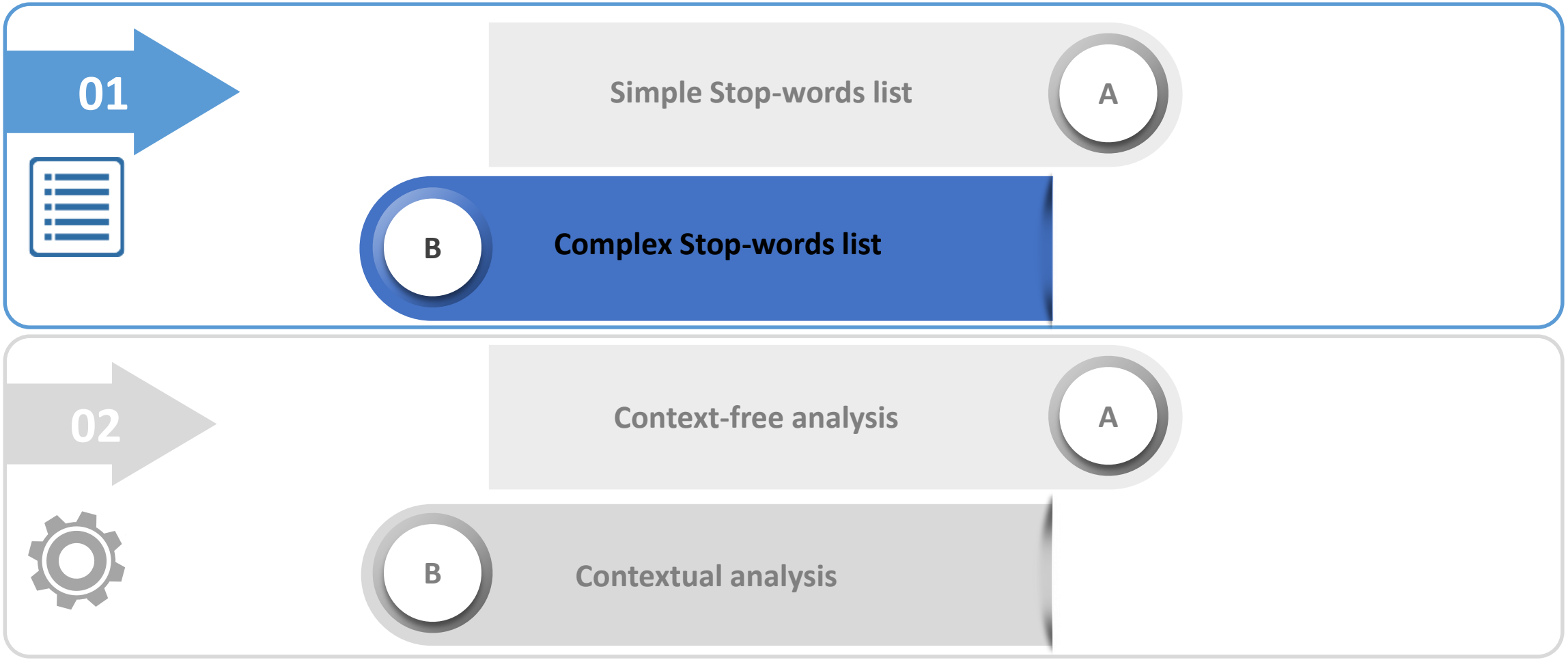
Proposed approach



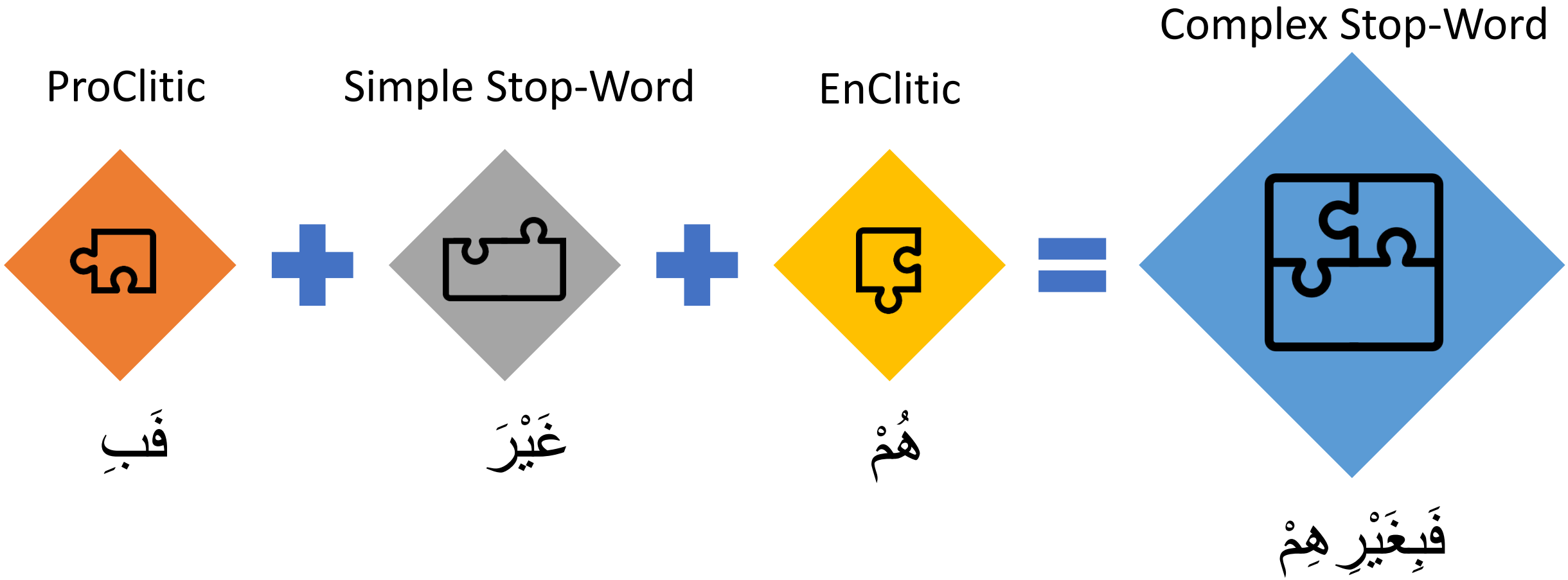
1.A- Simple Stop-words list design



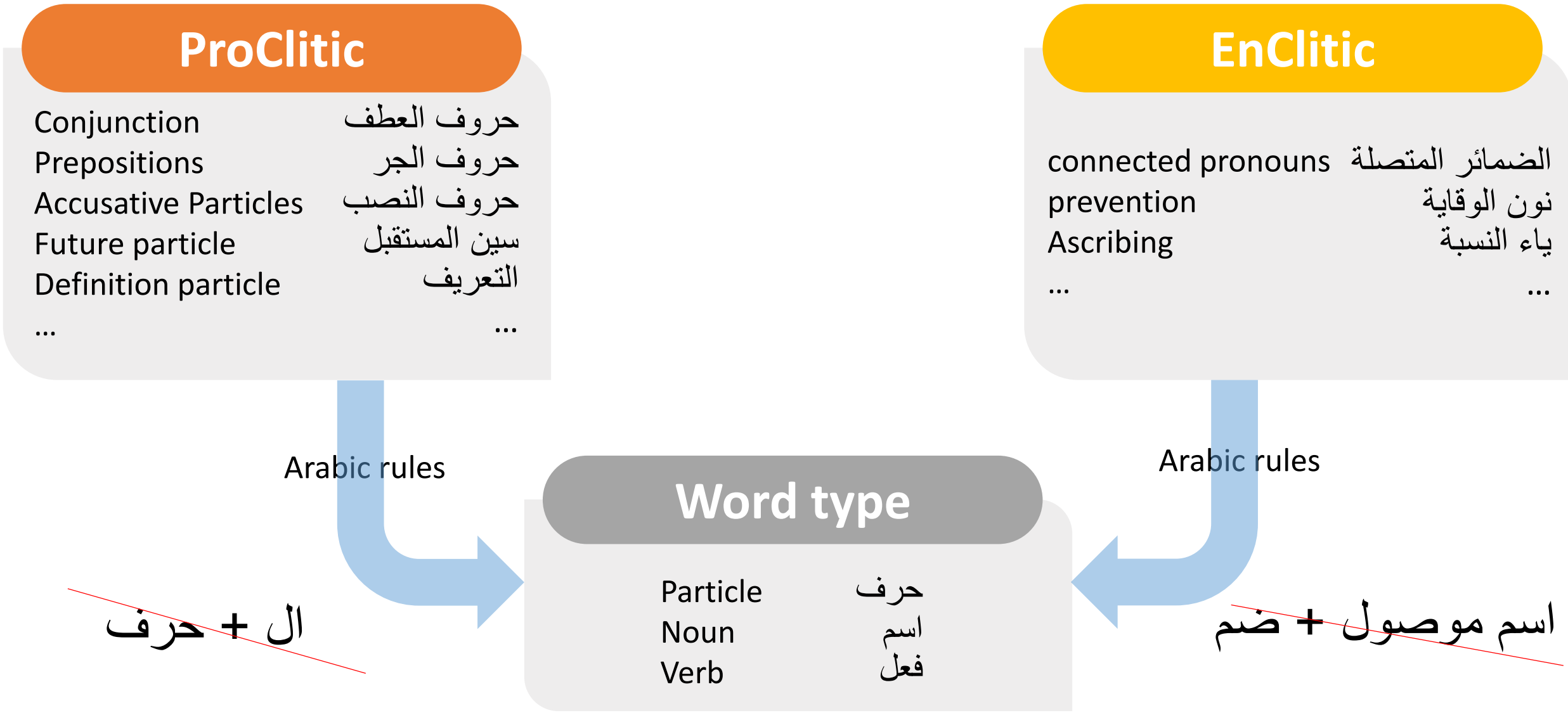
Proposed approach



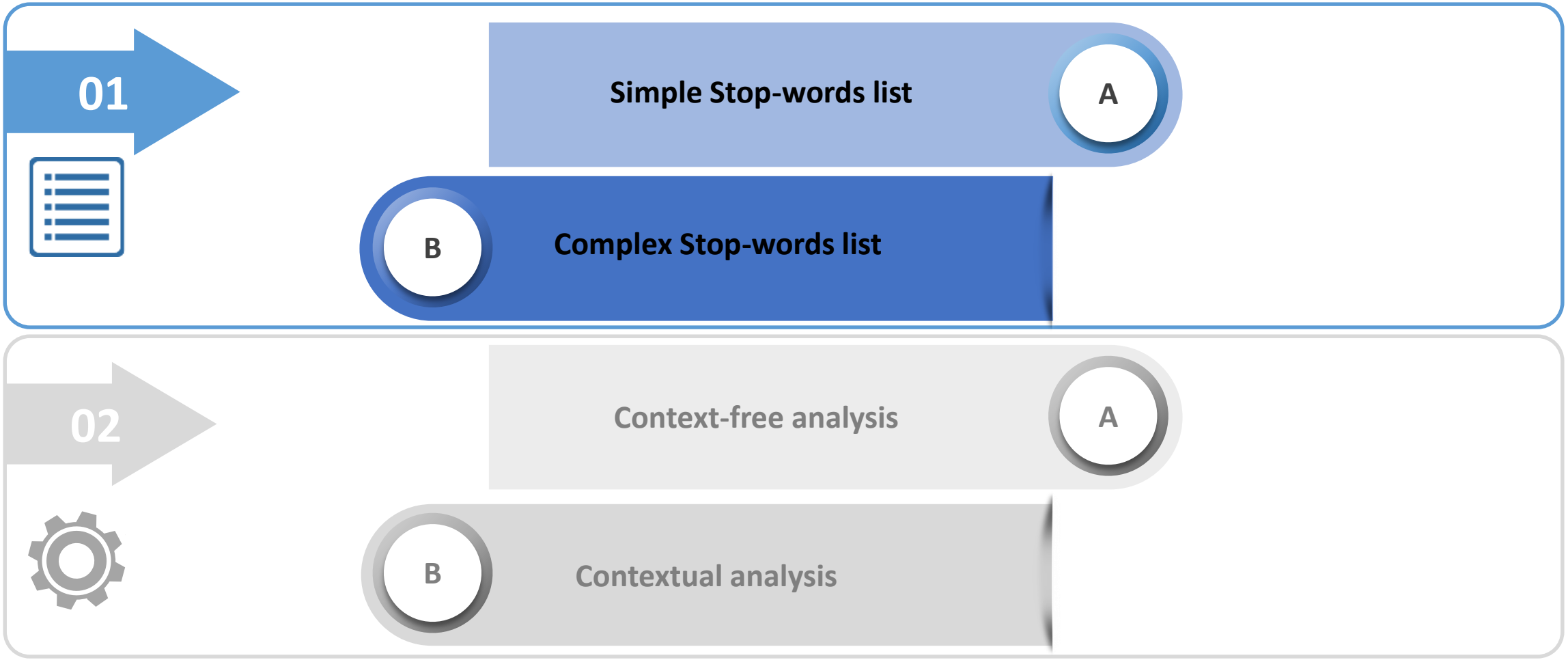
1.B- Complex Stop-words list design



1.B- Complex Stop-words list design



Proposed approach



1- Arabic Stop-words List (ASL)

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

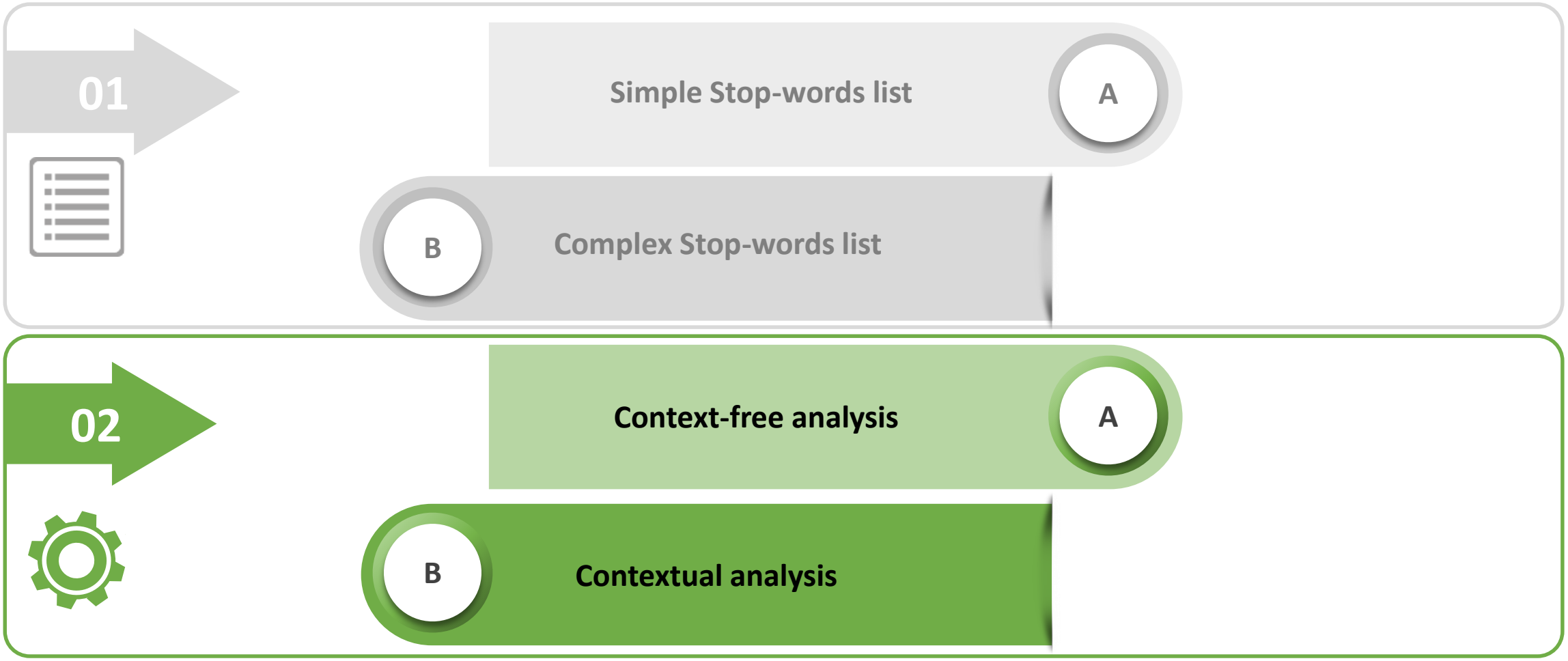
Enclitic	Simple stop-word		Proclitic	Complex stop-word		Category
-	maybe	عَلَّ	And وَلَ	And maybe	وَلَعَلَّ	NP
-	that	ذَلِكَ	like كَ	like that	كَذَلِكَ	SN
-	you are	تَكُونِينَ	will سَ	we will be	سَتَكُونِينَ	SV
you كِ	no more than	غَيْرُ	-	no more than you	غَيْرُكِ	SN
you كَ	maybe	عَسَى	-	Maybe you	عَسَاكَ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

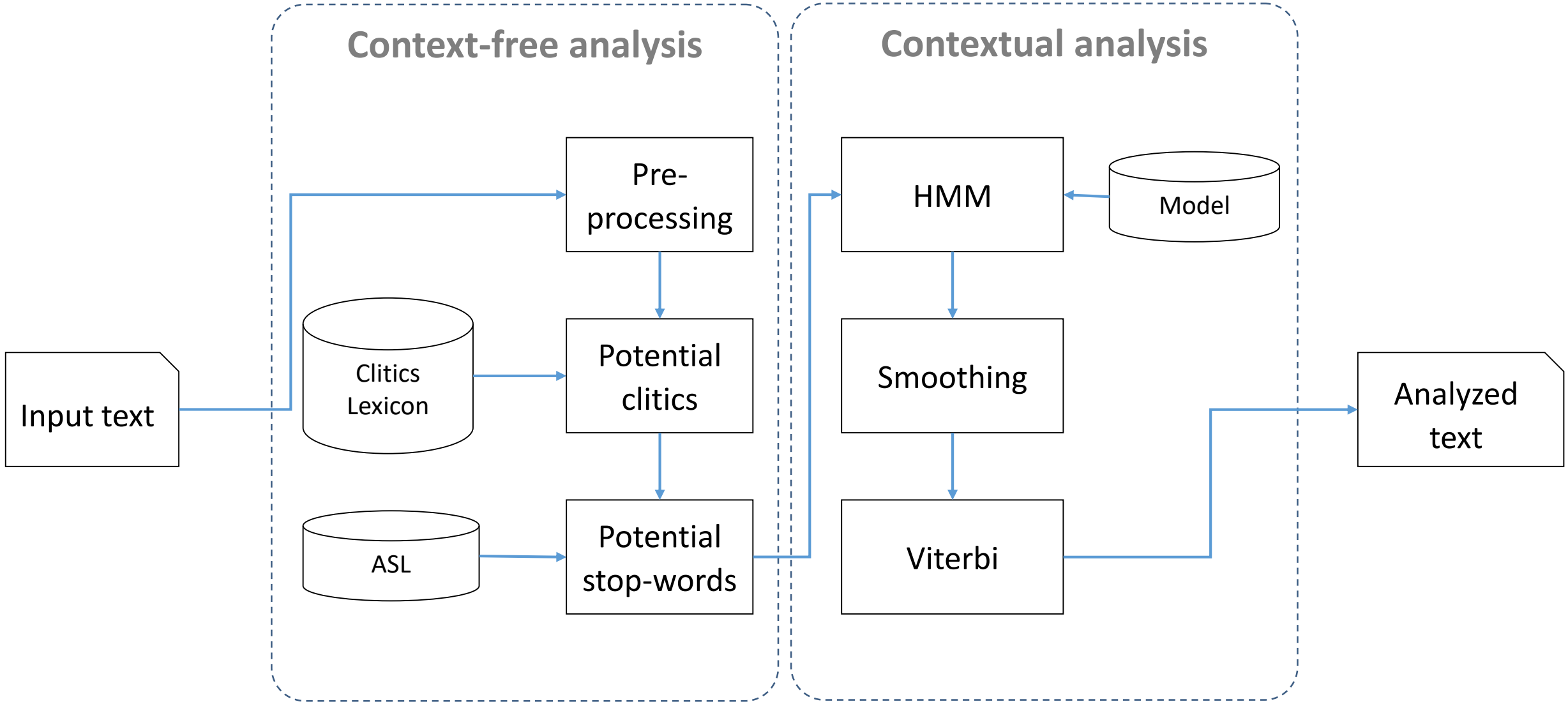
كُنْتَ، كُنَّا، كُنْتُمَا، كَانَتْ، كَانُوا، نَكُونُ، ...

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

Proposed approach

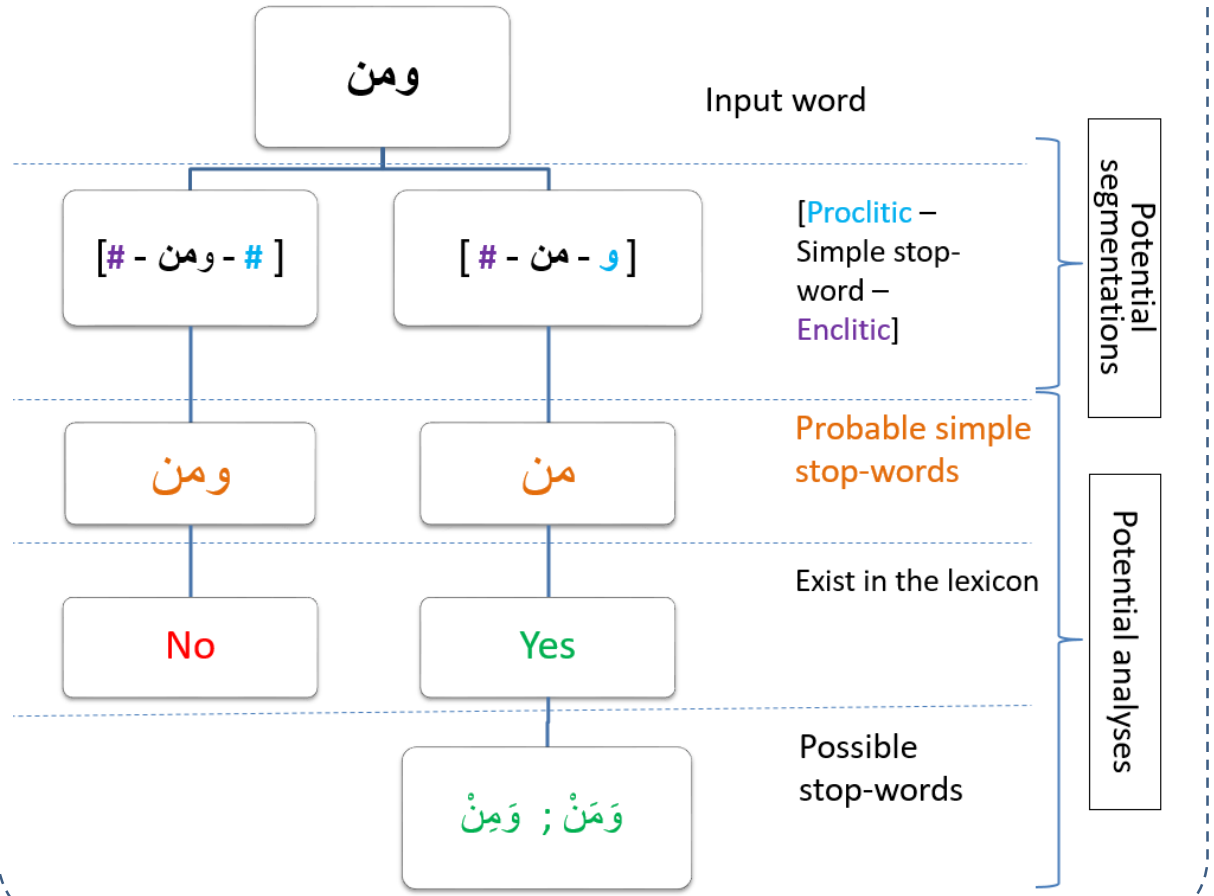


2- Stop-words analyzer design

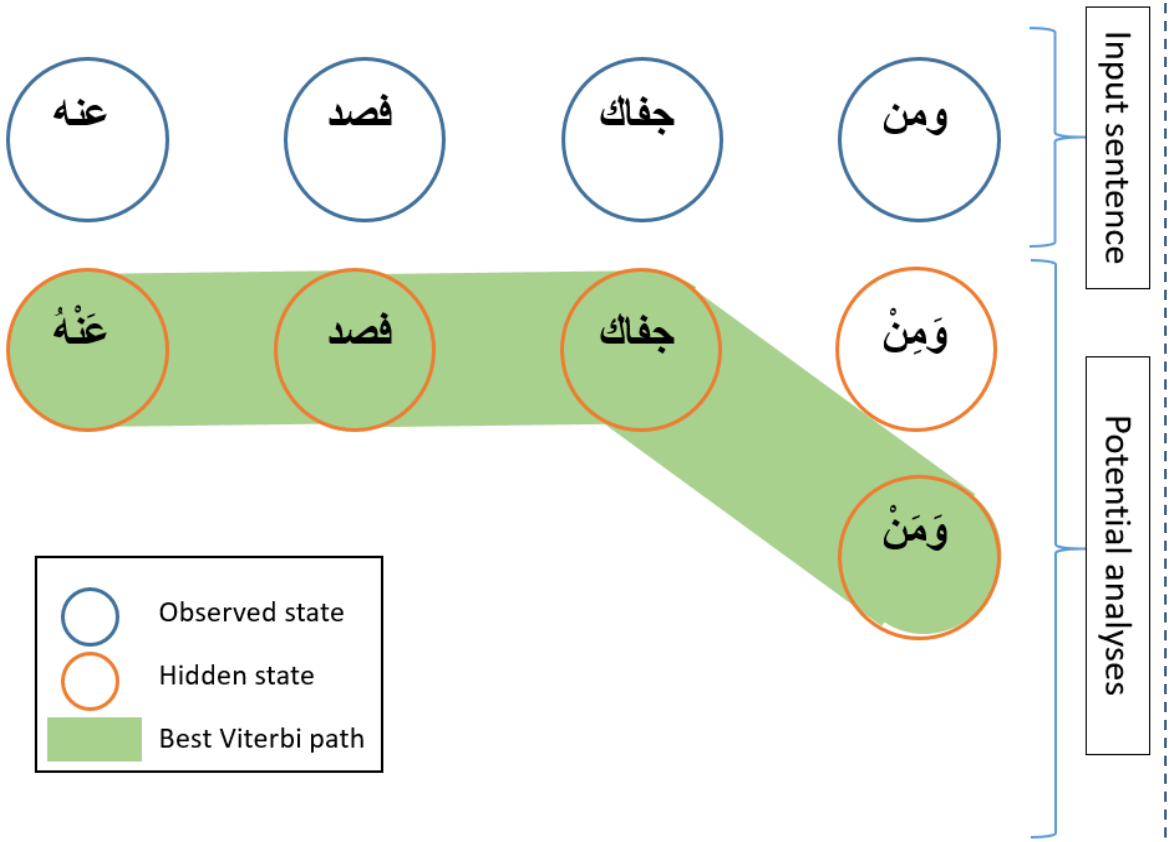


2- Stop-words analyzer design

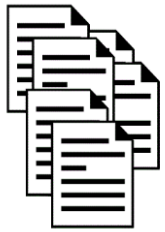
Context-free analysis



Contextual analysis



Stop-words list evaluation

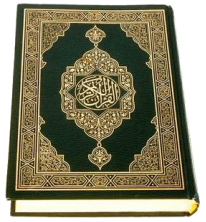


Tagged evaluation corpus

Compute

Accuracy

Exist



78,477 tokens

15,684 unique

هو	1
زيارة	0
…	…

Total
68

Khoja	Abu El-Khair	List 1	Ranks NL	ASL
144	552	98	68	1,247

Context-free evaluation

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

Accuracy

94. 11%

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

تنفرد المدينة العتيقة بمؤهلات متنوعة [وَلا] محدودة، انصهرت [فِيها] روافد ثقافية متباينة ميزتها [عَنّ، عُنْ، عَنْ] [غِيَرَها، غَيَّرَها] [مَنَّ، مُنْ، مِنْ، مَنٌّ، مَنْ] المدن

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

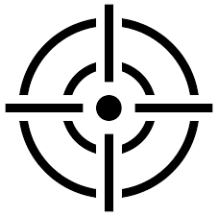
Contextual evaluation

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

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

تنفرد المدينة العتيقة بمؤهلات متنوعة [وَلا] محدودة،
انصهرت [فِيها] روافد ثقافية متباينة ميزتها [عَنّ، عُنْ، عَنْ]
[غِيَرَها، غَيَّرَها] [مَنَّ، مُنْ، مِنْ، مَنٌّ، مَنْ] المدن

Accuracy



82.35%

MADAMIRA	ASL analyzer
90.17	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%

Université Mohammed V de Rabat

ECOLE MOHAMMADIA D'INGENIEURS

Thank you

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

Driss Namly, Karim Bouzoubaa, Rachida Tajmout, Ali Laadimi

ALELM Team

ICALP 2019

16-17 Octobre- Nancy