# جامعة بوليتكنك فلسطين


# توسيم النصوص العربية بشكل آلي ودورها في معالجة اللغة العربية آليا

د. ضياء أبوزينة  د. محمود الصاحب

2014

# توسيم النصوص العربية

3. **وَسَّمَ:** ( فعل )

**وسَّمَ** يُوسِّم ، توسيمًا ، فهو **مُوسِّم** ، والمفعول **مُوسَّم**

**وسَّم** فلانًا : أعطاه أو منحه وسامًا

4. **وَسم:** ( إسم )

الجمع : **وُسُومٌ**

**الوَسْمُ** : سِمَة ، علامة

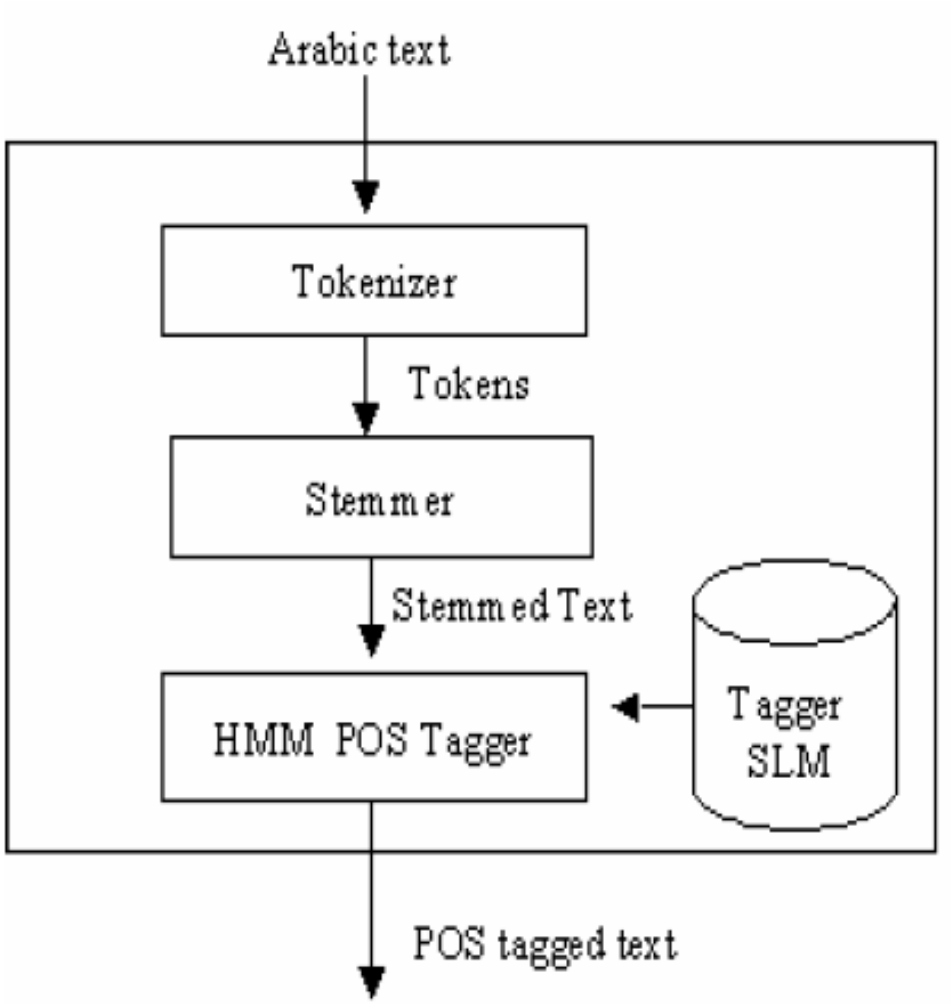
مصدر **وسَمَ**

**وَسْمُ** دَابَّةٍ : أَثَرُ الكَيّ

5. **وَسْم:** ( إسم )

**وَسْم** : مصدر **وَسَمَ**

# مكونات نظام توسيم الكلمات



# أهمية توسيم النصوص العربية

- عمليات التحليل اللغوي (Parsing)
- استرجاع المعلومات (Information Retrieval)
- التعرف على الكلام وتأليفة (Speech Recognition)
- انظمة المحاورة - الاسئلة والاجوبة (Question Answering systems)
- التنقيب عن البيانات (Data Mining)
- أنظمة المساعدة عبر الهاتف (Phone Help systems)
- تصنيف وتلخيص المقالات
- الترجمة الآلية
- التدقيق الإملائي
- كما يمكن الانتفاع منها في هندسة البرمجيات، والتطبيقات الدينية

# أهمية توسيم النصوص العربية

**حسب كتاب *التراكيب الشائعة في اللغة العربية – دراسة احصائية* (1982)**

**للدكتور محمد علي الخولي**

**"يفيد اللغة في كثير من الامور منها معرفة التراكيب الصرفية الشائعة في الاستعمال والتراكيب النادرة في الاستخدام. كما أن معرفة التراكيب الشائعة للغة العربية يساهم في عمليات التحليل لمقارنة اللغة المكتوبة باللغة المسموعة، ومقارنة اللغة الفصحى باللغة العامية، ومقارنة لغة النثر بلغة الشعر، ومقارنة اللغة الحديثة باللغة القديمة. هذا بالاضافة الى أن معرفة التراكيب الشائعة يفيد في تخطيط وتدريج المواد القرائية للمراحل الدراسية المختلفة سواء في كتب القراءة او كتب المواد الدراسية الاخرى"**

# توسيم النصوص بشكل عام

- إن مسألة توسيم الكلمات مسألة متشعبة اذ يختلف الباحثون في عددها لنفس اللغة
- لا يوجد مجموعة قياسية ثابتة في اللغات المختلفة
- اللغة العربية لها خصوصية فيما يتعلق بأقسام الكلمة الواحدة وبما تحتويه من ملحقات
- اشهر مجموعة وسوم ( Tag Set) مستخدمة في اللغة الانجليزية هي مجموعة (Benn Treebank) جامعة بنسلفانيا في الولايات المتحدة الامريكية وبها (45،36) وسم
- من اشهر المدونات للغة الانجليزية مدونة براون وتحتوي على مليون كلمة و (87 وسم)
- مجموعة (BNC) البريطانية بها 61 وسم
- نسبة الوسم الصحيح تتفاوت حسب المصادر المستخدمة وتتركز حول 95%
- احدى المشاكل الكبيرة في عملية وسم الكلمات هي الغموض (Ambiguity)

# أمثلة على توسيم النصوص

**اللغة الانجليزية**

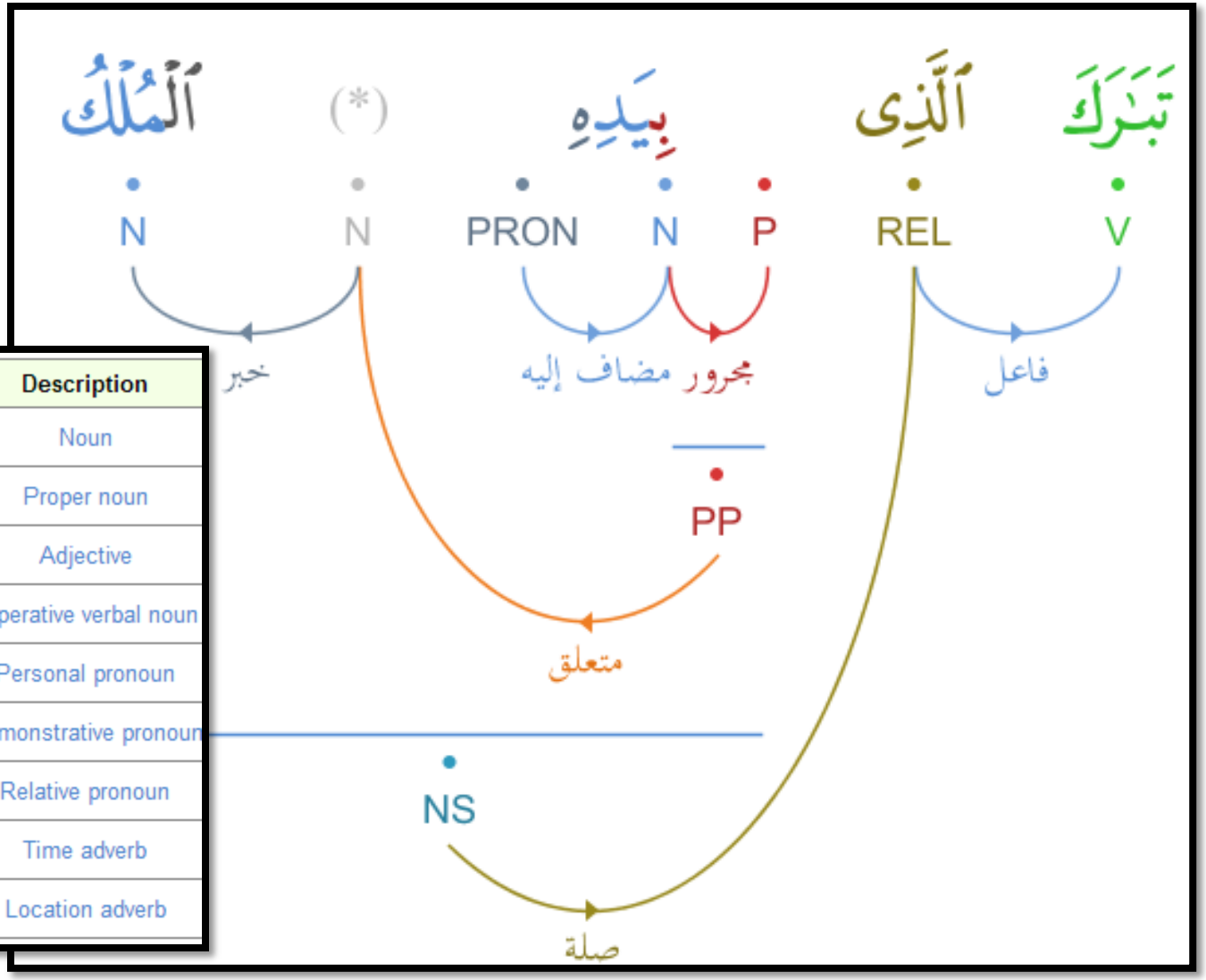
John/NNP saw/VBD the/DT saw/NN and/CC decided/VBD ...

**اللغة العربية**

قالت/VBD شركة/NN أرامكو/NNP السعودية/DTNNP
وشركة/NN دال/NNP كيميكلز/NNP الأمريكية/DTJJ
اليوم/DTNN

According to *Stanford* Tagger

# توسيم القرآن الكريم – جامعة ليدز




| Tag | Arabic Name | Description |
|---|---|---|
| N | اسم | Noun |
| PN | اسم علم | Proper noun |
| ADJ | صفة | Adjective |
| IMPN | اسم فعل أمر | Imperative verbal noun |
| PRON | ضمير | Personal pronoun |
| DEM | اسم اشارة | Demonstrative pronoun |
| REL | اسم موصول | Relative pronoun |
| T | ظرف زمان | Time adverb |
| LOC | ظرف مكان | Location adverb |

# مجموعة جامعة ستانفورد للغة العربية

| # | Tag | Meaning with examples | # | Tag | Meaning with examples |
|---|---|---|---|---|---|
| 1 | DTJJ | DT + Adjective<br>النفطية، الجديد | 16 | PRP | Personal pronoun<br>هي، هو |
| 2 | DTJJR | Adjective, comparative<br>الكبرى، العليا | 17 | PRP$ | Possessive pronoun<br>هم |
| 3 | DTNN | DT + Noun, singular or mass<br>المنظمة، العاصمة | 18 | RB | Adverb<br>هناك، حيث |
| ... | ... | ... | ... | ... | ... |
| 13 | CC | Coordinating conjunction<br>ثم ، و | 28 | ADJ | Adjective, Numeric<br>السابع، الرابعة |
| 14 | CD | Cardinal number<br>مئة، ألفين | 29 | UH | Interjection<br>unusual kind of word<br>(example: From English: oops) |
| 15 | DT | Demonstrative pronouns<br>هذه، ذلك | | | |

# من الاعمال في مجال وسم اللغة العربية

## A Hidden Markov Model –Based POS Tagger for Arabic

55 tags

Fatma Al Shamsi, Ahmed Guessoum

97%

| | | | | |
|---|---|---|---|---|
| ADJ | EXCEPT | PPRON_2FP | PRON_3D | SUFF_M_P |
| CONJ | FUNC_WORD | PPRON_3FP | PRON_3FP | SUFF_SUBJ_1P |
| CVERB | FUTURE | PREP | PRON_3FS | SUFF_SUBJ_2D |
| DEF | INTERROGATE | PRON | PRON_3MP | SUFF_SUBJ_2FP |
| DPRON_F | IV1P | PRON_1P | PRON_3MS | SUFF_SUBJ_2MP |
| DPRON_FD | IV2 | PRON_1S | PVERB | SUFF_SUBJ_2S |
| DPRON_FP | IV3 | PRON_2 | SHORT_FORM | SUFF_SUBJ_3FD |
| DPRON_FS | IVERB | PRON_2D | SUFF_F_D | SUFF_SUBJ_ALL |
| DPRON_MD | NEGATION | PRON_2FP | SUFF_F_P | SUFF_SUBJ_FP |
| DPRON_MP | NOUN | PRON_2MP | SUFF_F_S | SUFF_SUBJ_MP |
| DPRON_MS | PNOUN | PRON_2S | SUFF_M_D | SUFF_S_INDEF |

# من الاعمال في مجال وسم اللغة العربية

**Khoja's Tag set**

| | |
|---|---|
| Tag set size | 177 tags (103 types of noun, 57 verbs, 9 particles, 7 residuals,1 punctuation) |

**Penn Arabic Treebank (PATB) Part-of-Speech Tag Set (FULL)**

| | |
|---|---|
| Tag set size | The FULL tag set comprises over 2000 tag types. This includes combinations of 114 basic tags. |

**Penn Arabic Treebank (PATB) Reduced Part-of-Speech Tag Set (RTS)**

| | |
|---|---|
| Tag set size | 25 tags |

**Penn Arabic Treebank (PATB) Extended Reduced Part-of-Speech Tag Set**

| | |
|---|---|
| Tag set size | 75 tags |

**ARBTAGS**

| | |
|---|---|
| Tag set size | 161 detailed tags (101 nouns, 50 verbs, 9 particles, 1 punctuation mark including 28 different POS general tags to cover the main part-of-speech classes and sub-classes. |

# طرق التوسيم

**اولاً: طريقة القواعد**

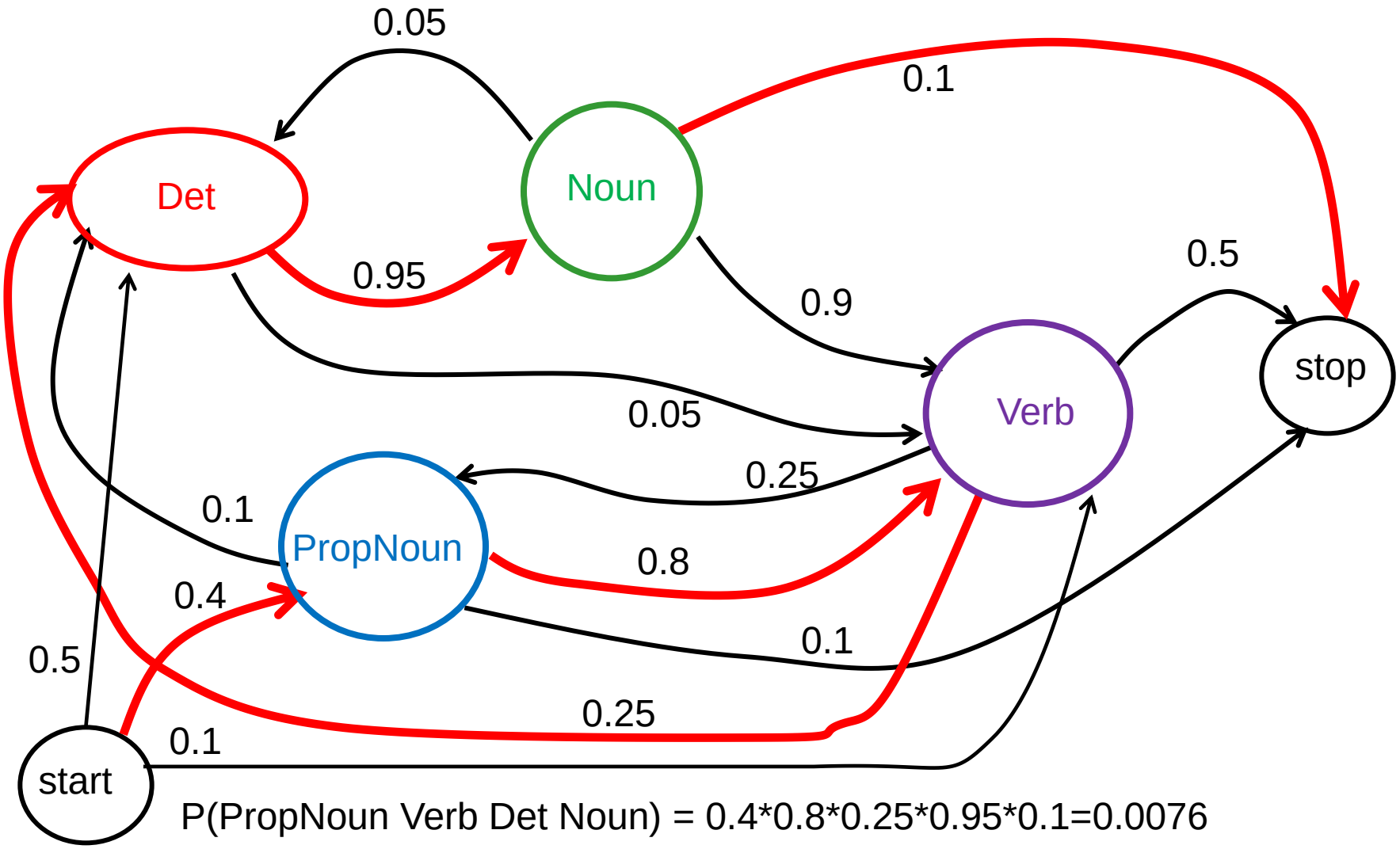
**يتم وضع القواعد من قبل اللغويين او يتم استنباطها من مدونة تم توسيمها مسبقاً**

**ثانيا: الطريقة الاحصائية**

**يتم توسيم الكلمات بالاعتماد على احتمالية ورودها حسب الوسوم المتجاورة وهذه الطريقة بحاجة الى التدريب (لبناء نماذج ماركوف المخفية، على سبيل المثال)**

**طريقة القواعد اكثر دقة إلا انها اكثر كلفة**

# مثال على نماذج ماركوف المخفية




P(PropNoun Verb Det Noun) = 0.4*0.8*0.25*0.95*0.1=0.0076

# الصيغ الصرفية للغة العربية - الدكتور محمد الخولي

| الاسماء |
| --- |
| في حالة الجمع، في حالة التثنية، في حالة الافراد، المبني، المعرب، المذكر ،المؤنث، المحايد، المعرف بال،النكرة المضاف الى معرفة،الضمير،الاسم الموصول،اسم الاشارة، العلم، الاسم النكرة،الاسم المعرفة، المصدر،اسم المرة،اسم الهيئة،اسم الفاعل،الصفة المشبهة ،اسم المفعول ،افعل التفضيل،صيغة المبالغة،اسم المكان ،اسم الزمان،اسم الالة،ضمير المتكلم، ضمير المخاطب،ضمير الغائب ،الضمير المتصل،الضمير المنفصل، جمع المذكر السالم،جمع المؤنث السالم،جمع التكسير، الاسماء الخمسة، النسبة، التصغير،المقصور ،المنقوص، الممدود،الاسماء المرفوعة،الاسماء المنصوبة، الاسماء المجرورة،الفاعل المذكر،الفاعل المؤنث،نائب الفاعل،المبتدأ،خبر المبتدأ،ضمير الشأن، ضمير الفصل، اسم كان،اسم كاد،خبر أن،خبر لا النافية للجنس،التابع المرفوع، المفعول المطلق ،نائب المفعول المطلق،المفعول به المذكر، المفعول به المؤنث،المفعول به لعامل مذكر،المفعول به على التحذير، المفعول به على الاغراء، المفعول به على الاختصاص، المفعول به على الاستغاثة،المفعول به على الندبة، المنادى، المرخم، المفعول لاجله ،ظرف الزمان، ظرف لمكان،المفعول معه ،الحال، التمييز،خبر كان واخواتها،اسم ان واخواتها،اسم لا النافية للجنس، التابع المنصوب،المضاف اليه، المجرور بحرف الجر،التابع المجرور، البدل ،الصفة،التوكيد اللفظي، التوكيد المعنوي،المعطوف |

# الصيغ الصرفية للغة العربية - الدكتور محمد الخولي

| الافعال |
| --- |
| الفعل الناقص، الفعل التام،الفعل الماضي،الفعل المضارع،فعل الامر،الفعل اللازم،الفعل المتعدي،الفعل المعلوم،الفعل المجهول،الفعل ذو الفعل الظاهر،<br>الفعل ذو الفعل المستتر، الفعل المتعدي لمفعول واحد، الفعل المتعدي لمفعولين،<br>الفعل المتعدي لثلاثة مفاعيل، الفعل المضارع المنصوب،الفعل المضارع لمجزوم،<br>الفعل المضارع المرفوع،الفعل الثلاثي المجرد،<br>الفعل الرباعي المجرد،الفعل الرباعي المزيد،<br>الفعل الصحيح، الفعل المعتل،كان واخواتها،كاد واخواتها،افعال المدح والذم،<br>الفعل المؤكد بنون التوكيد،اسم الفعل، |

# الصيغ الصرفية للغة العربية - الدكتور محمد الخولي

| الحروف |
| --- |
| حروف الجر: في ، بِـ ، من ، لِـ، لـَ ، على ، الى ، عن ، كـَ ، حتى |
| حروف القسم |
| حروف العطف:( و ، فـَ ، أو ، ثم ، بل ، أم) |
| ان واخواتها: (أنّ ، لكن ، ليت ، لعلّ ، كأنّ ) لا النافية،قد |
| حروف نصب المضارع: (أنْ ، لام التعليل ، حتى ، فاء السببية ، واو المعية ، لام الجحود ، لن ، كي ، أذن) |
| حروف جزم المضارع: (لم ، لما ، لام الامر ، لا الناهية، إنْ ، اذ ، ما) |
| ربما ، أن المصدرية غير الناصبة |
| تاء التانيث |
| نون الافعال الخمسة |
| نون الوقاية |
| حروف التسويف |
| حروف اخرى |

# دقة أنظمة وسم الكلمات

$$\text{Accuricy percentage} = \frac{\#\text{of correctly tages words}}{\#\text{of words in the evaluation set}} * 100$$

- يوجد معايير اخرى للقياس
- الدقة تعتمد على عوامل متعددة منها عدد الوسم وحجم البيانات

# خلاصة الموضوع

**لا بد من ايجاد مجموعة وسوم موحدة معيارية او ايجاد الية لربط المجموعات مع بعضها البعض حتى يتم الانتفاع بها. وبغير ذلك فإن كل باحث سيبدأ بمجموعته الخاصة ولن تتحقق الفائدة**

- **تم مراجعة عدد من مجموعات الوسوم**
- **جاري العمل على ايجاد رابط بين المجموعات**

# THANK YOU

## Questions?