المدقق اللغوي النحوي

مفتوح المصدر للغة العربية

طه زروقي جامعة البويرة، الجزائر
صهيب عفيفي، جامعة أرتوا، فرنسا

التدقيق اللغوي النحوي

تصميم مشروع مدقق لغوي للغة للعربية

- للنحو والتركيب
- الأسلوب
- مفتوح المصدر

أنواع التدقيق

التدقيق الإملائي: يهتم بالكلمة دون سياقها

التدقيق النحوي: يهتم بالكلمة في سياقها

إنشاء الله



أنواع التدقيق

التدقيق الإملائي: يهتم بالكلمة دون سياقها

التدقيق النحوي: يهتم بالكلمة في سياقها

إنشاء الله

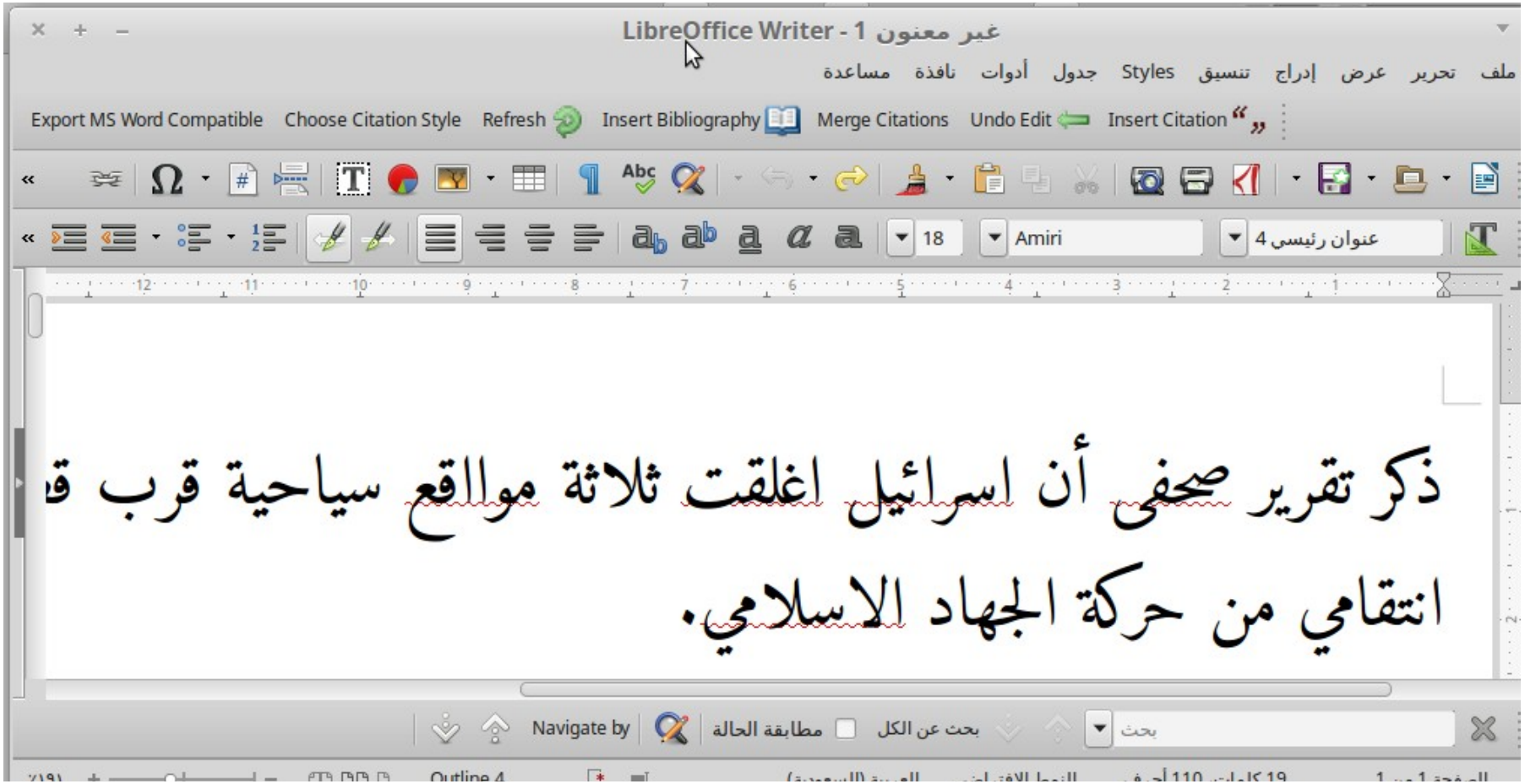
التدقيق الإملائي

التدقيق الإملائي

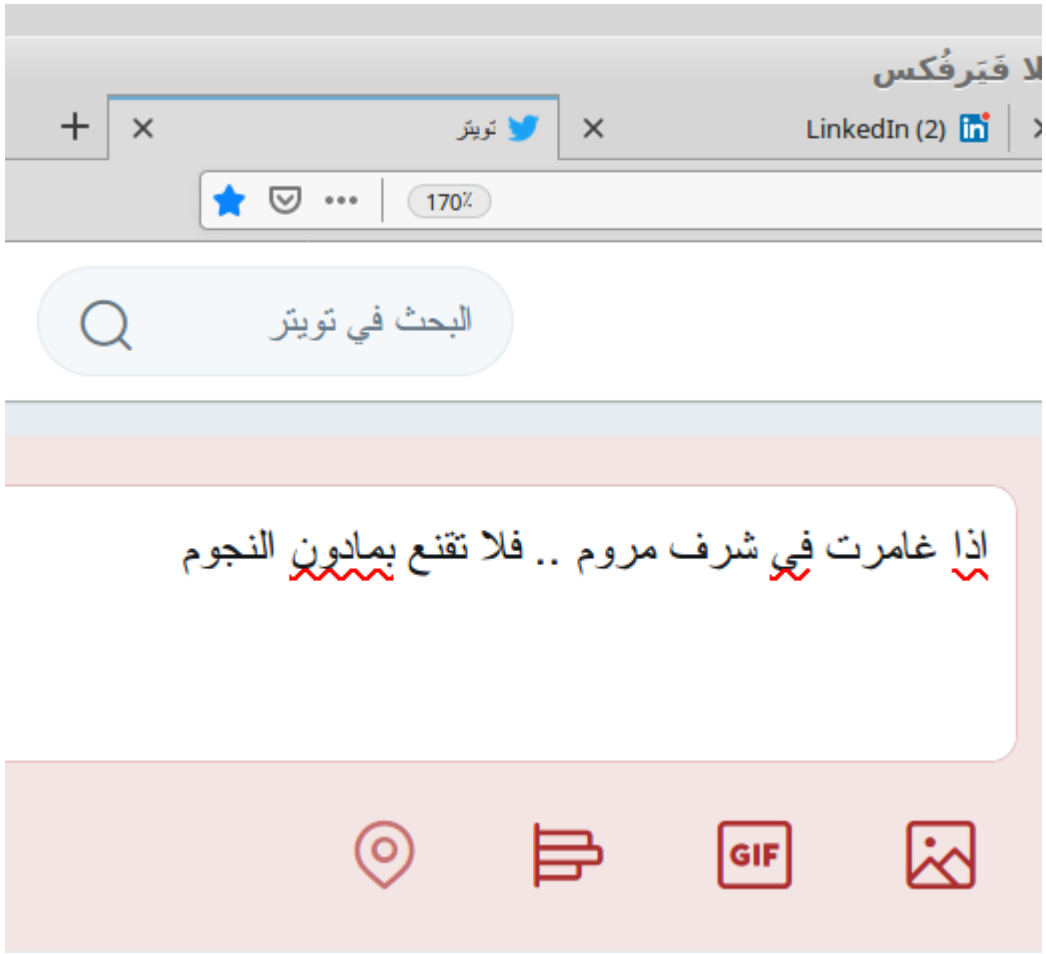
- مشروع آية سبل
  - مدقق عربي حر
  - متعدد المنصات والبرامج
  - مستخدم في برامج التصفح والمكتبية.
- مشروع غلطاوي للتصحيح التلقائي

LibreOffice®

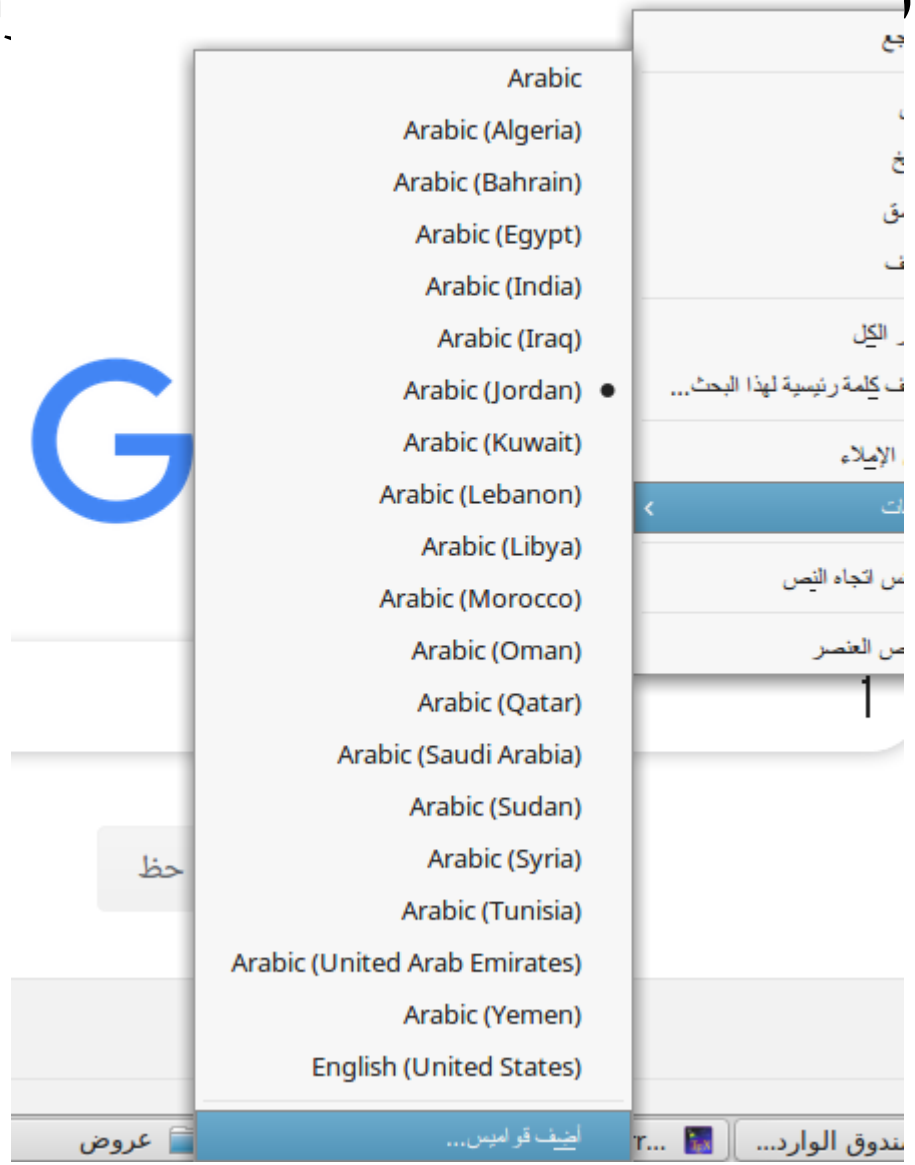
تجربة على المكتب



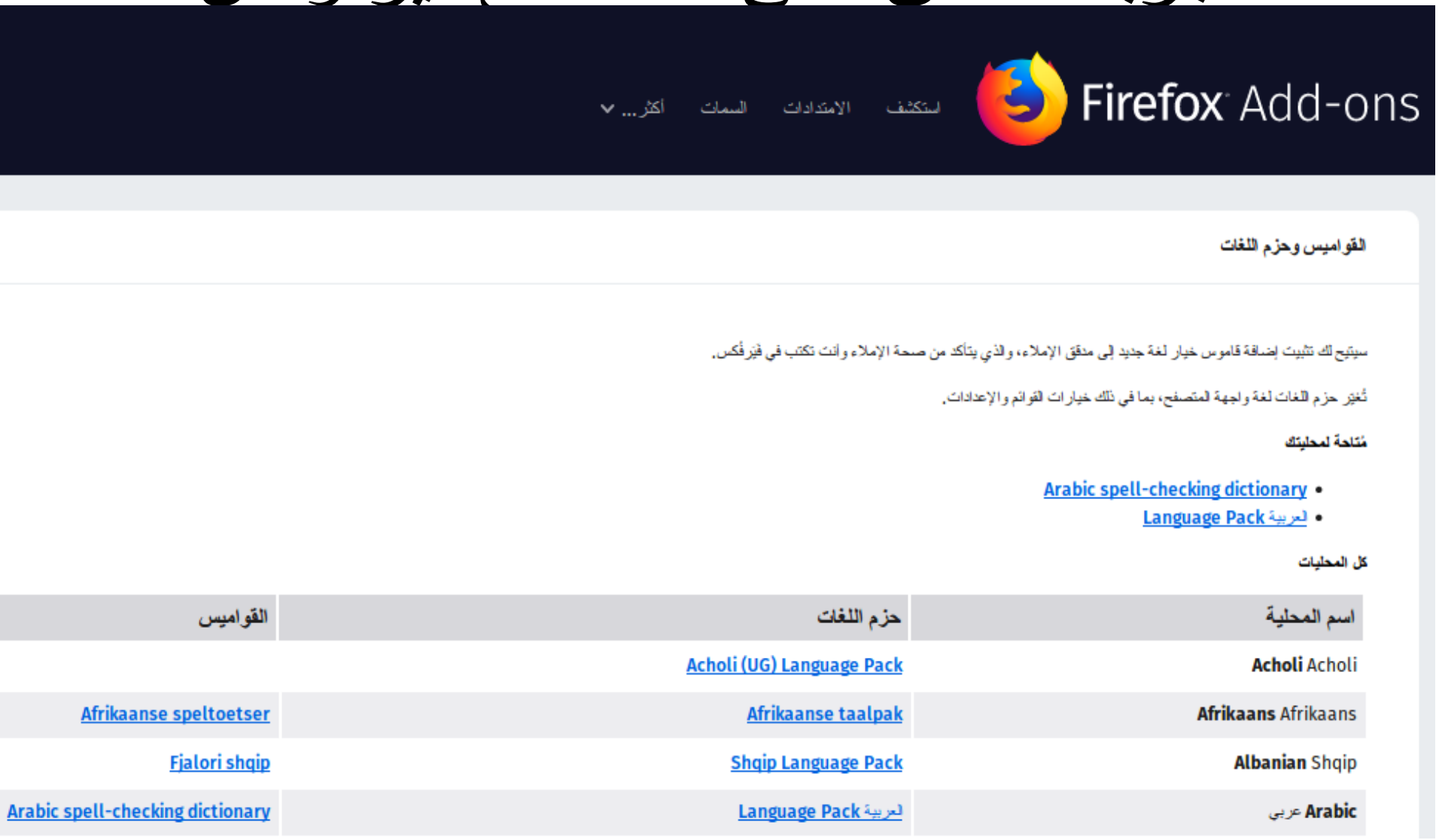
تجربة المدقق على المتصفح فيرفوكس



تجربة ا فوكس


Arabic
Arabic (Algeria)
Arabic (Bahrain)
Arabic (Egypt)
Arabic (India)
Arabic (Iraq)
Arabic (Jordan)
Arabic (Kuwait)
Arabic (Lebanon)
Arabic (Libya)
Arabic (Morocco)
Arabic (Oman)
Arabic (Qatar)
Arabic (Saudi Arabia)
Arabic (Sudan)
Arabic (Syria)
Arabic (Tunisia)
Arabic (United Arab Emirates)
Arabic (Yemen)
English (United States)
عروض

تجربة المدقق على المتصفح فيرفوكس





آية سبل









التصحيح التلقائي

الأخطاء الإملائية الشائعة

يرتبك العربي في كتابة همزة الوصل والقطع في كلمة

*إستعمال

فعلى الرغم من صحة الكلمة في باقي حروفها، إلا أنها تعاني من خطإ نمطي، وتصحح فقط باستبدال همزة القطع بهمزة الوصل.

كيف يعمل

- قائمة الكلمات
- التعبيرات المنتظمة

قائمة الكلمات

اسلام	إسلام
ضلام	ظلام
المكتبه	المكتبة
اعلام	إعلام

التعبيرات المنتظمة

قواعد الوزن انفعال

(و|ف|)(ك|ب|)(ال|)إن(\w\w)ا(\w)ي|)(ين|ات|ة|تين|)

وبالإنفعالات

التعابير المنتظمة

- قائمة الكلمات
- التعابير المنتظمة

التصحيح التلقائي

اذا أردت إستعارة كتاب، اذهب الى المكتبه او الادارة فى الضهيرة

إذا أردت استعارة كتاب، اذهب إلى المكتبة أو الإدارة في الظهيرة

/http://ghalatawi.sourceforge.net



تطبيقات

- قائمة التصحيح التلقائي في برامج المكتبية OpenOffice/ LibreOffice
- تدقيق مواقع ذات المحتوى الضخم، مثل ويكيبيديا.
- مساعدة الكاتب في برامج المراسلات، والترجمة، وغيرها.
- تدقيق المدوّنات اللغوية (corpus).
- تصحيح استعلامات البحث.
- مساعدة الكتابة في الأجهزة المحمولة كالهواتف الذكية.

التدقيق النحوي

التدقيق اللغوي

التدقيق اللغوي يهتم بالكلمة في سياقها يشمل

- النحو
- الأسلوب
- الدلالة

التدقيق اللغوي

التدقيق اللغوي يهتم بالكلمة في سياقها يشمل

- النحو
- الأسلوب
- الدلالة

إلى المواطنون

إلى المواطنين

لم يجري

لم يجر-

التدقيق اللغوي

التدقيق اللغوي يهتم بالكلمة في سياقها يشمل

- النحو
- الأسلوب
- الدلالة

القيام بعملية الضرب

يضرب

علاوة على ذلك

إضافة إلى ذلك

التدقيق اللغوي

التدقيق اللغوي يهتم بالكلمة في سياقها يشمل

- النحو
- الأسلوب
- الدلالة

كسر القانون

خالف

رجل بسيط

غرّ، مغفل، ساذج

أنواع الأخطاء

يمكن تصنيف الأخطاء إلى

- التطابق
- الاستعمال

كسر القانون

خالف

رجل بسيط

غرّ، مغفل، ساذج

التطابق

يمكن تصنيف الأخطاء إلى

- في العدد
- في النوع

خمسة طالبات

خمس طالبات

هذا بئر

هذه بئر، وبئر معطلة وقصر مشيد

الاستعمال

يمكن تصنيف الأخطاء إلى

- التعدي بحرف
- الحركات
- تركيب غير مناسب

يحتاجه

يحتاج إليه

تمادى على

تمادى في

حِراك

حَراك- فتح الحاء-

الاستعمال

يمكن تصنيف الأخطاء إلى

- التعدي بحرف
- الحركات
- تركيب غير مناسب
- جمع التكسير

باب مغلوق

باب مغلق

مشاكل

مشكلات

وقف مذهولا

وقف ذاهلا

مشروع التدقيق

- الأعمال المشابهة
- مراحل المشروع
- تقدّم المشروع

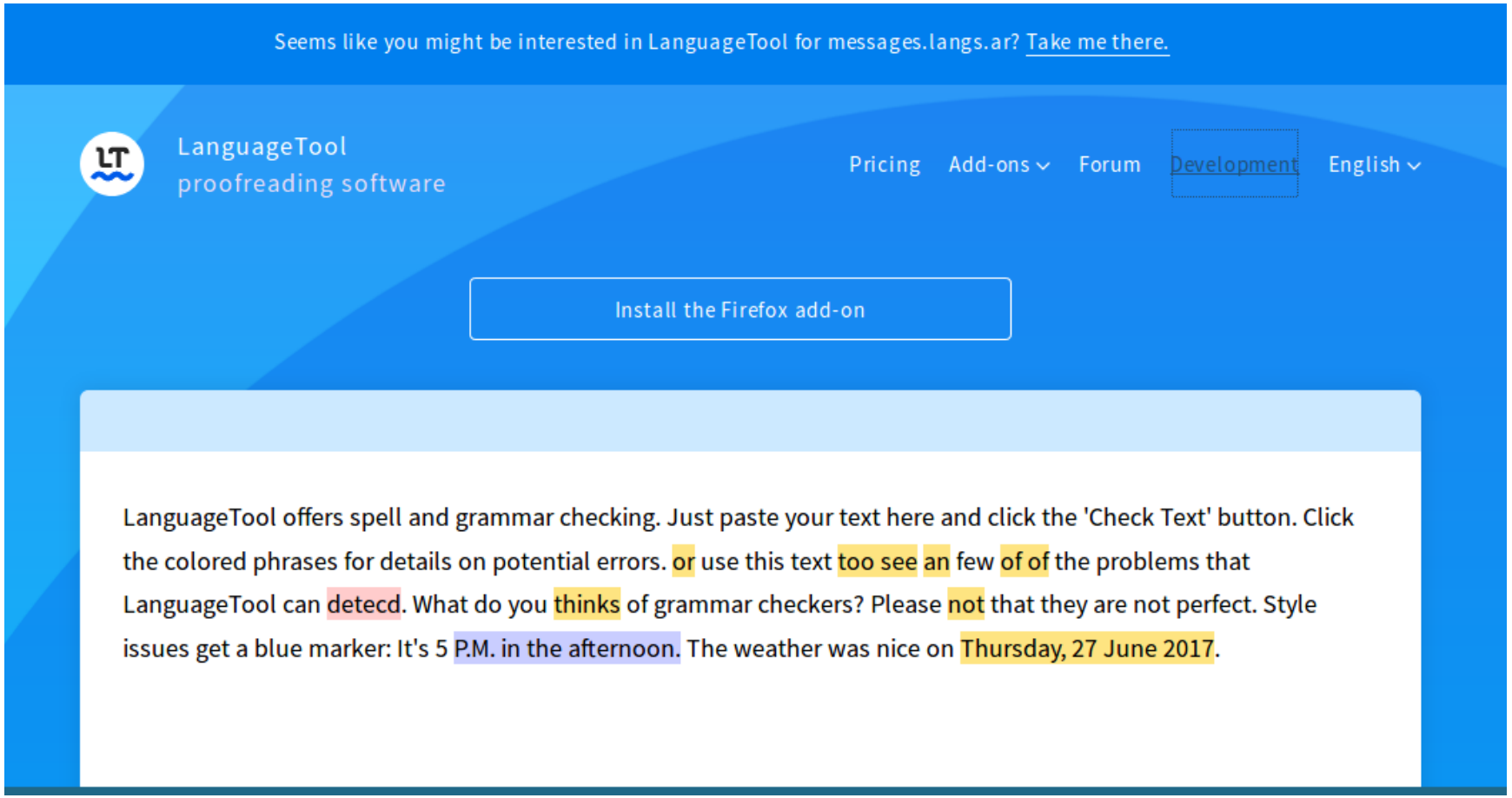
الأعمال المشابهة

Language Tool	Arabic GramCheck	Microsoft Word	Light Proof
مفتوح المصدر	مغلق	مغلق	مفتوح المصدر
قيد التنفيذ	2005	2003	2011
متوفر	غير متاح	متوفر تجاريا	متوفر

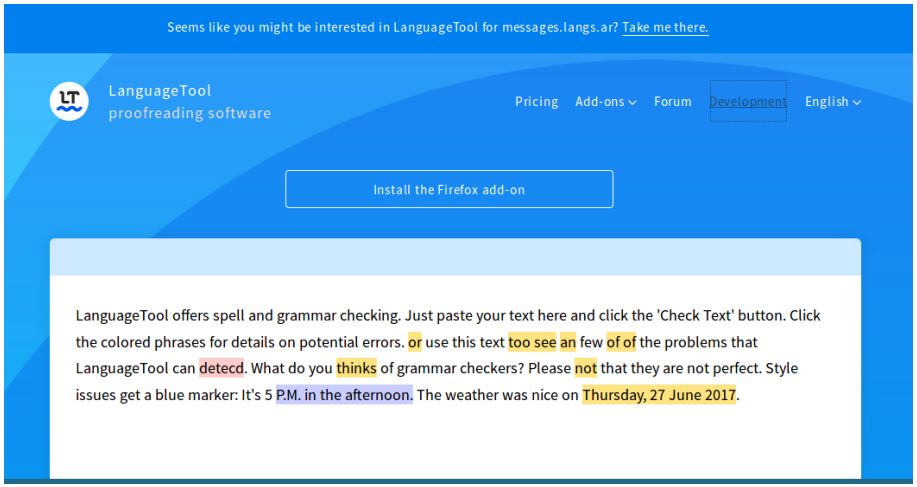
برنامج أداة اللغة LanguageTool

- مدقق لغوي نحوي أسلوبي
- مفتوح المصدر
- متعدد اللغات

Languagetool‚org



برنامج أداة اللغة LanguageTool



- يوفر عدة أشكال
- موقع وب
- تطبيق مكتبي
- إضافة للمكتبية والمتصفح

LibreOffice
The Document Foundation

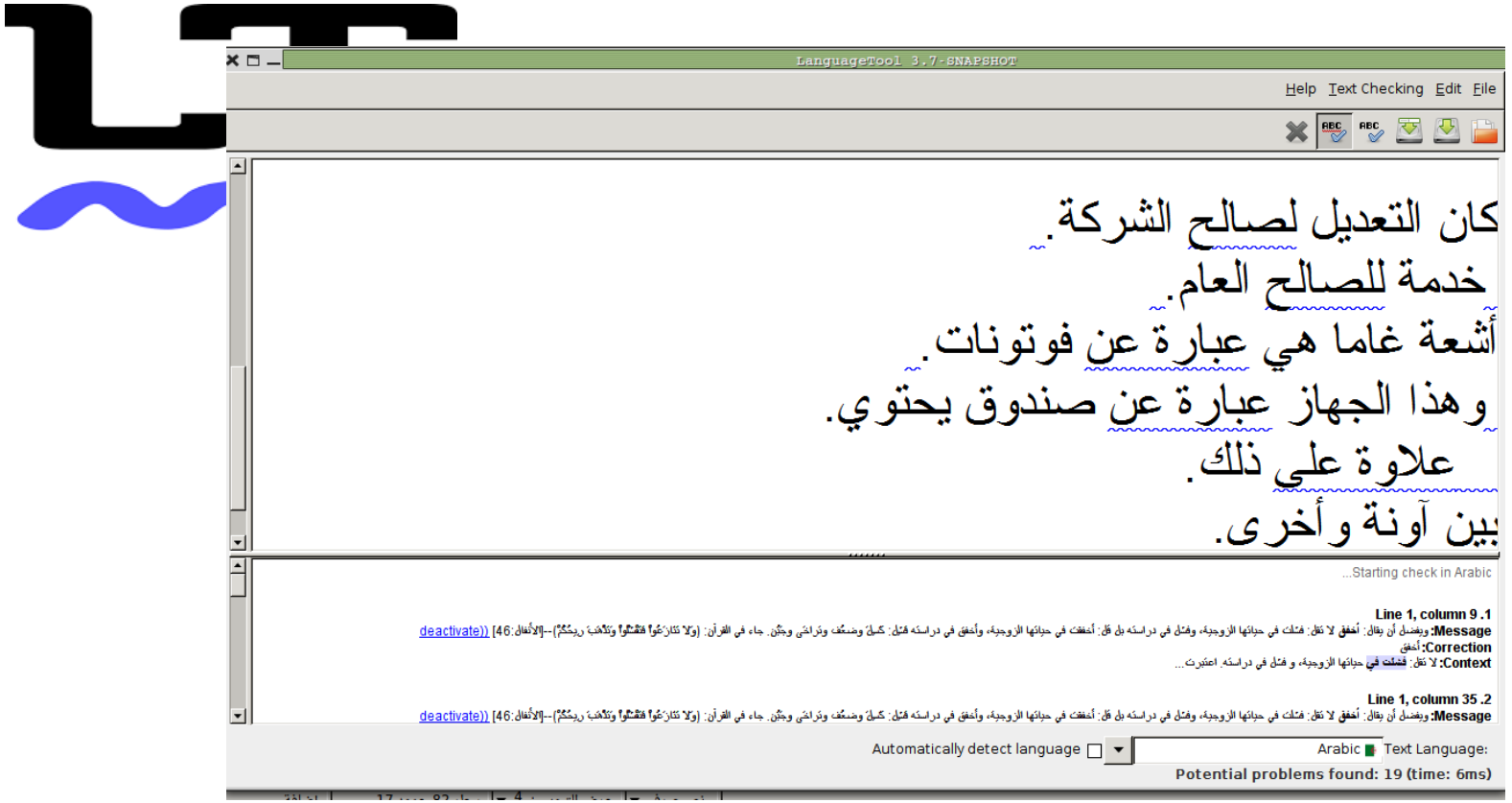


LanguageTool

الجانب البرمجي

- لغة البرمجة جافا
- مقسّم إلى وحدات
- تمثيل القواعد بلغة XML
- المرونة
- API ولجهة برمجية

التدقيق اللغوي النحوي



دعم العربية في أداة اللغة

مشروع دعم العربية في أداة اللغة ذو ثلاث أقسام

- بناء وحدة البرمجية للعربية
- بناء قاموس قواعد كشف الأخطاء
- ترجمة الواجهات

دعم العربية في أداة اللغة

مشروع دعم العربية في أداة اللغة ذو ثلاث أقسام

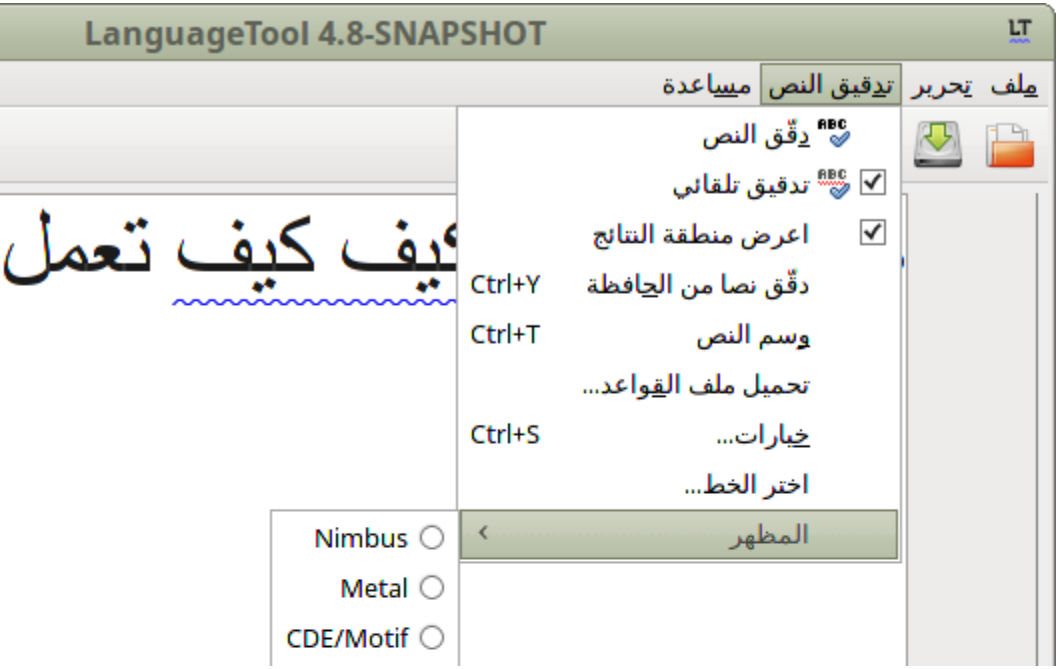
- بناء وحدة البرمجية للعربية
- بناء قاموس قواعد كشف الأخطاء
- ترجمة الواجهات

- برمجة مقسّم الكلمات
- برمجة آلية الوسم

دعم العربية في أداة اللغة

مشروع دعم العربية في أداة اللغة ذو ثلاث أقسام

- بناء وحدة البرمجية للعربية
- بناء قاموس قواعد كشف الأخطاء
- ترجمة الواجهات



دعم العربية في أداة اللغة

مشروع دعم العربية في أداة اللغة ذو ثلاث أقسام

- بناء وحدة البرمجية للعربية
- بناء قاموس قواعد كشف الأخطاء
- ترجمة الواجهات

- نمذجة قواعد كشف الأخطاء
- بناء قاموس الوسوم

قاموس الوسوم

يوفر النظام آلية لبناء قاموس للكلمات مع وسومها،
كل كلمة نضع لها ما يقابلها من معلومات على شكل

وباللغتين
لغة: اسم جامد، مؤنث، مثنى، معرفة، مجرور بالباء معطوف بالواو

وباللغتين[لغة/WBL;-F2A;-NJ

ترميز الوسم -الاسم-

الوسم ذو ثلاثة أقسام للاسم للفعل

القسم الأول: القاموس

- نوع الكلمة: اسم، فعل، حرف
- نوعها الفرعي: جامد، اسم فاعل، صفة

القسم الثاني: التصريف

القسم الثالث: الزوائد

NJ-;F2A-;WBL

ترميز الوسم

الوسم ذو ثلاثة أقسام للاسم للفعل

القسم الأول: القاموس

القسم الثاني:التصريف

- النوع: مذكر، مؤنث
- العدد: 1،2،3
- الحالة الإعرابية: مرفوع، منصوب، مجرور
- العلامة الإعرابية: حركة، حرف

القسم الثالث:الزوائد

NJ-;F2A-;WBL

ترميز الوسم

الوسم ذو ثلاثة أقسام للاسم للفعل

القسم الأول: القاموس

القسم الثاني:التصريف

NJ-;F2A-;WBL

القسم الثالث:

- معطوف بالواو بالفاء، غير معطوف
- مجرور بالباء/الكاف/اللام، غير مجرور
- معرف بال، نكرة، ضمير متصل.

ترميز الوسم - الفعل-

ويستعملونها

اِسْتَعْمَلَ

V-1;M3H-faU;W-H

ترميز الوسم - الفعل-

الوسم ذو ثلاثة أقسام للفعل

القسم الأول: القاموس

V-1;M3H-faU;W-H

– نوع الكلمة: اسم، فعل، حرف

– نوعها الفرعي: معتل، سالم

– متعدي/لازم، متعدي لمفعولين

القسم الثاني:التصريف

القسم الثالث: الزوائد

ترميز الوسم - الفعل-

الوسم ذو ثلاثة أقسام للفعل

القسم الأول: القاموس

V-1;M3H-faU;W-H

القسم الثاني:التصريف

- نوع الضمير: مذكر، مؤنث
- العدد: 1،2،3
- الشخص: متكلم، مخاطب، غائب
- العلامة:حركة، حرف
- الزمن: ماض مضارع، أمر
- البناء: معلوم، مجهول
- الحالة: مرفوع، منصوب، مجزوم

القسم الثالث: الزوائد

ترميز الوسم - الفعل-

الوسم ذو ثلاثة أقسام للفعل

القسم الأول: القاموس

القسم الثاني:التصريف

القسم الثالث: الزوائد

- معطوف بالواو بالفاء، غير معطوف
- حرف الاستقبال
- ضمير متصل.

V-1;M3H-faU;W-H

مزايا طريقة الوسم

يتميز هذا الوسم بأنّ كل حرف في موضع معين يمثل حالة معينة، ويمكن تمثيل غياب الخاصية أيضا.

يمكن تمثيله بعبارة نمطية regular expression

عيوب:

صعب القراءة.

V.*;M3.*;---

V-1;M3H-faU;W-H

مزايا طريقة الوسم

يمكن تمثيله بعبارة نمطية regular expression

الأسماء المؤنثة المفردة	`N.*;F1.*`
الأسماء الجمع	`N.*;.?3.*`
الأسماء المجرورة بالباء	`N.*;.*;-B.?`

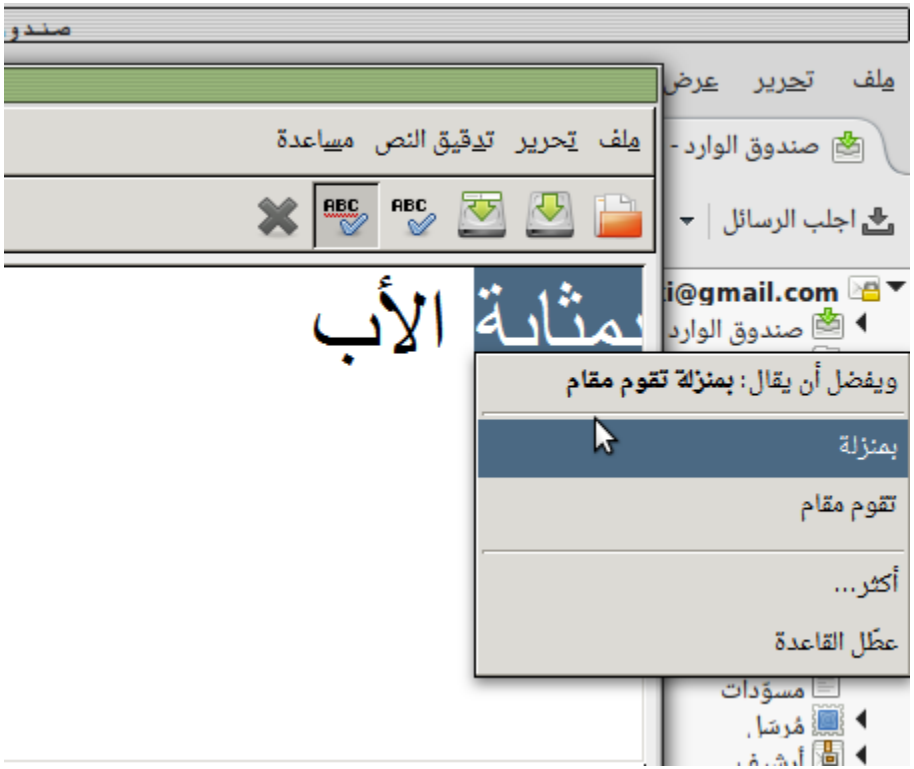
مزايا طريقة الوسم

العيوب

صعبة القراءة دون الرجوع للدليل

V-1;M3H-faU;W-H

التدقيق اللغوي النحوي



كتابة قواعد كشف الأخطاء

أداة اللغة تضع صياغة بلغة XML لكتابة قواعد كشف الأخطاء

كتابة قواعد كشف الأخطاء

```
<rule>
  <pattern>
      <marker>  <token> هذا  </token>  </marker>
      <token regexp="yes">الكتف|الفخذ|الساق|الريح|البئر</token>
  </pattern>
  <message>الاسم مؤنث.. أتقصد
    <suggestion>هذه</suggestion>
  </message>

  <example correction="هذه"><marker>هذا</marker> الكتف/example>
</rule>
```

كتابة قواعد كشف الأخطاء

```
<rule>
  <pattern>
      <marker>  <token> اهذ  </token>  </marker>
      <token regexp="yes">البئر|الريح|الساق|الفخذ|الكتف</token>
  </pattern>
  <message>تقصد أ..مؤنث الاسم
    <suggestion>هذه</suggestion>
  </message>
  <example correction="هذه"><marker>اهذ</marker> الكتف/example>
</rule>
```

النمط

الرسالة

الاقتراحات

الأمثلة

كتابة قواعد كشف الأخطاء

```
<rule>
<pattern>
        <marker>  <token> هذا  </token>  </marker>
        <token>البئر</token>
 </pattern>
```

النمط يحوي على نسق من الوحدات النصية TOKEN

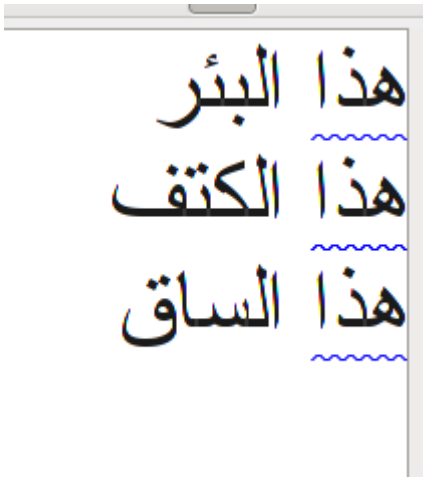
- الوحدة النصية هي كلمة متصلة
- ويمكن أن تكون تعبيرا نمطيا REGEX
- كلمة في كل حالاتها الصرفية
- أو وسما معبرا عن حالة معينة للكلمة

كتابة قواعد كشف الأخطاء

```
<rule>
<pattern>
        <marker>  <token> هذا  </token>  </marker>
        <token regexp="yes">البئر|الريح|الساق|الفخذ|الكتف</token>
 </pattern>
```

النمط يحوي على نسق من الوحدات النصية TOKEN

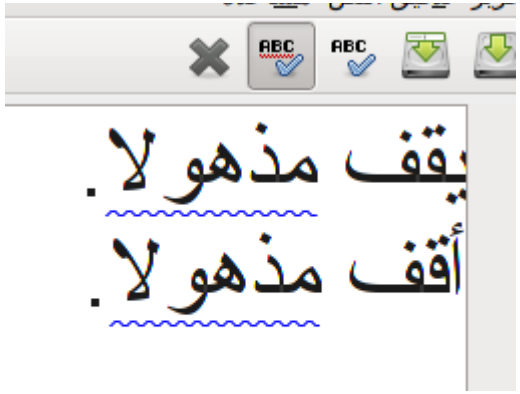
- الوحدة النصية هي كلمة متصلة
- ويمكن أن تكون تعبيرا نمطيا REGEX
- كلمة في كل حالاتها الصرفية
- أو وسما معبرا عن حالة معينة للكلمة



كتابة قواعد كشف الأخطاء

```
<rule>
<pattern>
    <marker> <token inflected="yes”> وَ قَفَ</token> </marker>
  <token > مذ هو لاـ</token>
 </pattern>
```

النمط يحوي على نسق من الوحدات النصية TOKEN



- الوحدة النصية هي كلمة متصلة
- ويمكن أن تكون تعبيرا نمطيا REGEX
- كلمة في كل حالاتها الصرفية
- أو وسما معبرا عن حالة معينة للكلمة
- هذا + كل الأسماء المؤنثة المفردة

كتابة قواعد كشف الأخطاء

```
<rule>
<pattern>
    <marker> <token> هذا </token> </marker>
    <token postag="N.*;F1.*;--.*" postag_regexp="yes"/>
</pattern>
```

النمط يحوي على نسق من الوحدات النصية TOKEN

- الوحدة النصية هي كلمة متصلة
- ويمكن أن تكون تعبيرا نمطيا REGEX
- كلمة في كل حالاتها الصرفية
- أو وسما معبرا عن حالة معينة للكلمة
  - هذا + كل الأسماء المؤنثة المفردة

كتابة قواعد كشف الأخطاء

```
<rule>
  <pattern>
      <marker>  <token> اهذ  </token>  </marker>
      <token regexp="yes">البئر|الريح|الساق|الفخذ|الكتف</token>
  </pattern>
  <message>تقصد أ..مؤنث الاسم
    <suggestion>هذه</suggestion>
  </message>
  <example correction="هذه"><marker>اهذ</marker> الكتف/example>
</rule>
```

النمط

الرسالة

الاقتراحات

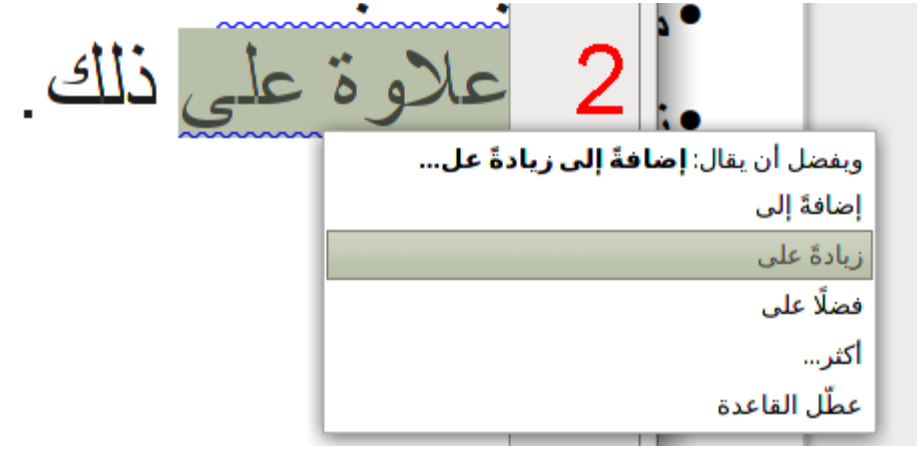
الأمثلة

كتابة قواعد كشف الأخطاء: الرسالة والاقتراحات

الاقتراحات يمكن أن تكون

- متعددة
- تعبير منتظم regex

<message> ويفضل أن يقال:
<suggestion> إضافةً
إلى <suggestion/>
<suggestion> زيادةً
على <suggestion/>
<suggestion> فضلًا على <suggestion/>
لأن العِلاوة: أعلى الرأس، أو ما يُحمل على البعير وغيره. وهي ليست مصدرًا.
<message/>

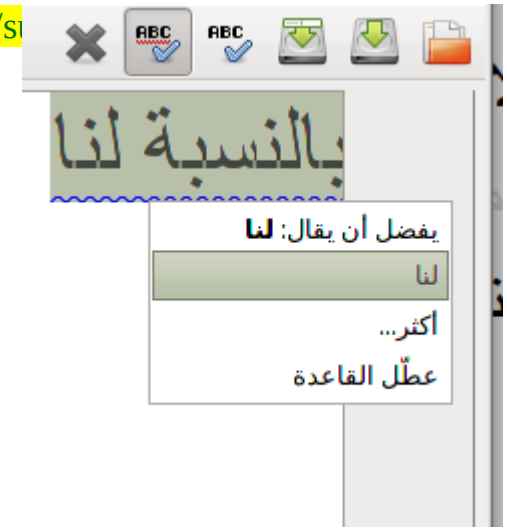


كتابة قواعد كشف الأخطاء: الرسالة والاقتراحات

الاقتراحات يمكن أن تكون

- متعددة
- تعبير منتظم regex
  - حذف الكلمة الأولى واقتراح الكلمة الأولى فقط.

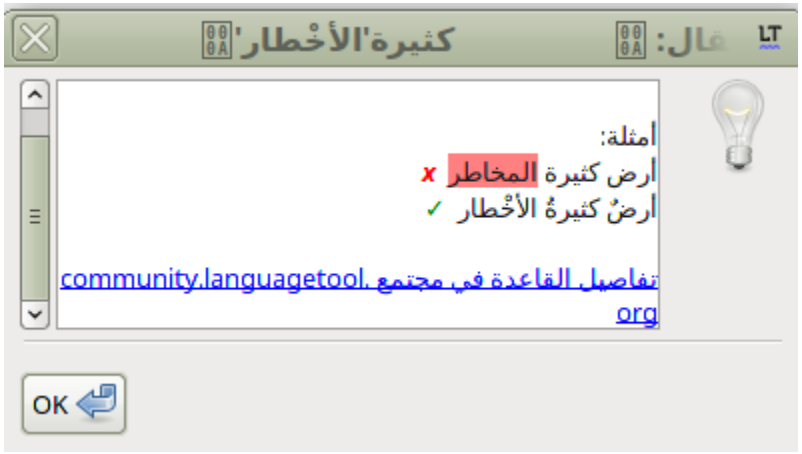
```
<pattern>
<marker>
<token/>بالنسبة<token>
<token/>"لي|لنا|لك"<"token regexp="yes>
<marker/>
<pattern/>
<message>يفضل أن يقال:
<suggestion><match no="2"/></su
<message/>
```



كتابة قواعد كشف الأخطاء: الأمثلة

```
<example correction='الأخْطار' type='incorrect'> أرض كثيرة
    <marker>المخاطر</marker> </example>
```

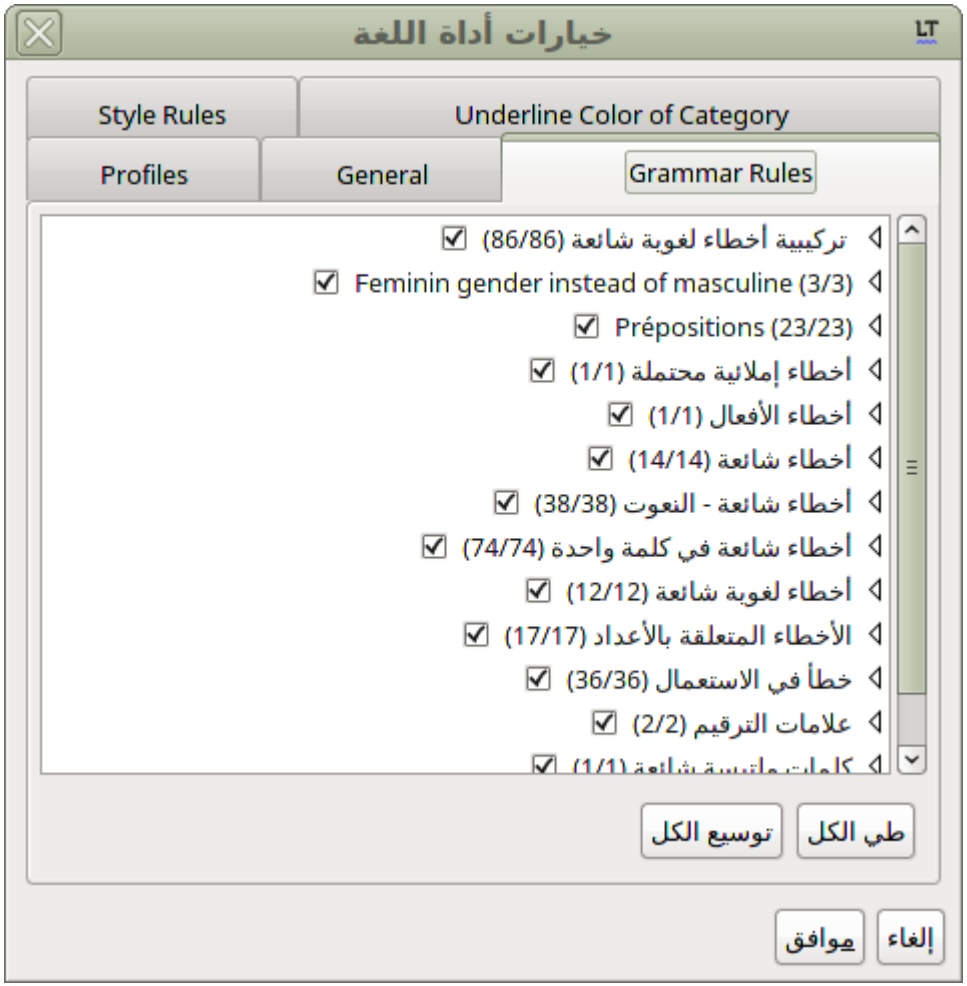
```
<example type='correct'> أرضٌ كثيرةُ الأخْطار </example>
```



ملف القواعد

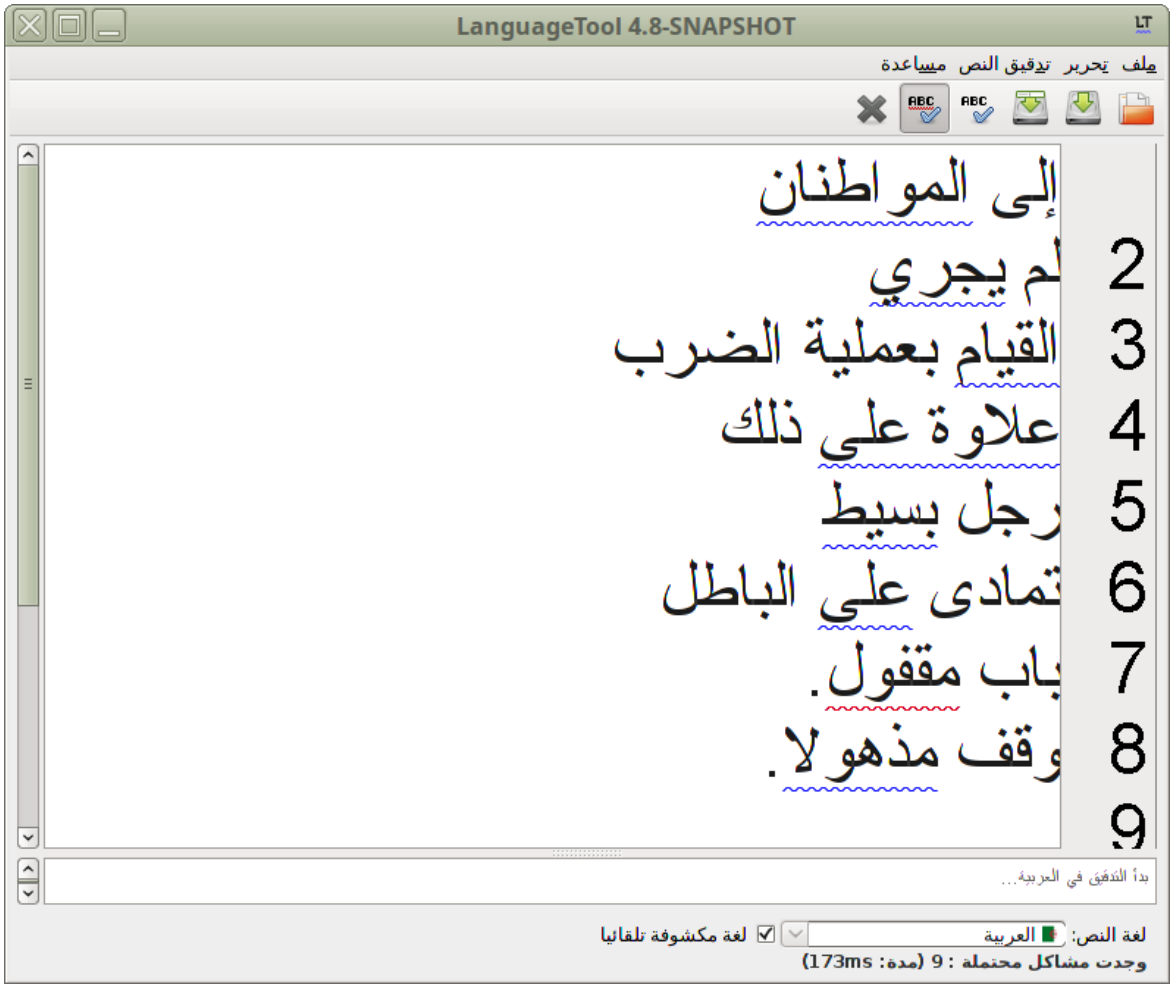
توضع كل القواعد بصيغة XML في ملف
Grammar.xml
داخل مجلد وحدة اللغة العربية.

ملف القواعد



قائمة بالقواعد المنجزة

مثال عام


LanguageTool 4.8-SNAPSHOT
ملف تحرير تدقيق النص مساعدة
إلى المواطنان
2 لم يجري
3 القيام بعملية الضرب
4 علاوة على ذلك
5 رجل بسيط
6 تمادى على الباطل
7 باب مقفول.
8 وقف مذهولا.
9
بدأ التدقيق في العربية...
لغة النص: العربية
لغة مكشوفة تلقائيا
وجدت مشاكل محتملة : 9 (مدة: 173ms)

الملفات الإضافية

coherency.txt

diacritics.txt

redundancies.txt

replaces.txt

wordiness.txt

wrongWordInContext.txt

Darja.txt

ملف التجانس

coherency.txt

أحيانا بعض الكلمات لها كتابات مختلفة كلها صحيحة:

مثلا:

شئون وشؤون

مائة و مئة

لذا يستحسن أن تكتب الكلمة بطريقة واحدة في نفس المستند، لذا فإنّ المدقق ينبهك إلى أنّك كتبت نفس الكلمة بطريقتين مختلفتين.

ملف التجانس

coherency.txt

أحيانا بعض الكلمات لها كتابات مختلفة كلها صحيحة:

مثلا:



لذا يستحسن أن تكتب الكلمة بطريقة واحدة في نفس المستند، لذا فإنّ المدقق ينبهك إلى أنّك كتبت نفس الكلمة بطريقتين مختلفتين.

ملف الحركات

diacritics.txt

أحيانا بعض الكلمات نطقها الشائع غير صحيح، لذا يستحسن التنبيه على حركتها، وهي ليست خاطئة بلا حركات:

مثلا:

حِراك



حَراك- فتح الحاء-

ملف العامية

darja.txt

في الكلمات العاميةوما يقابلها من الفصيح

طرشي=فلفل حلو

فايدة=فائدة

راديو=مذياع

سكانر=ماسح ضوئي

سبيطار=مستشفى

ملف المكررات

redundancies.txt

بعض العبارات فيها حشو أي زائد، يستحسن أن تصحح،

مثلا:

سوف لن يسير

لن يسير

هذه العبارات يمكن تمثيلها في ملف القواعد، لكن وضعها هنا يسهل إضافتها دون كتابة قاعدة.

ملف الاستبدال

replaces.txt

الملف يحوي جدولا للكلمات وبدائلها الصحيحة، يمكن إثراؤها دون كتابة قواعد جديدة.

مثلا:

رئيف=رؤوف

خارطة=المُصوَّر|خريطة

رفاة=رفات

يمكن وضع اقتراحات لكل كلمة.

ملف .....

wordiness.txt

أحيانا بعض الكلمات نطقها الشائع غير صحيح، لذا يستحسن التنبيه على حركتها، وهي ليست خاطئة بلا حركات:

مثلا:



طازج

طازَج- بفتح الزاي

ملف .....

wrongWordInContext.txt

أحيانا بعض الكلمات نطقها الشائع غير صحيح، لذا يستحسن التنبيه على حركتها، وهي ليست خاطئة بلا حركات:

مثلا:



طازج

طازَج- بفتح الزاي

الملفات الإضافية

coherency.txt

diacritics.txt

redundancies.txt

replaces.txt

wordiness.txt

wrongWordInContext.txt

الموارد اللسانية

- معجم أخطاء الكتاب، صلاح الدين الزعبلاوي.
- دليل الأخطاء الشائعة في الكتاب والنطق، د. مروان البواب.
- نحو إتقان الكتابة العلمية باللغة العربية، محمد مكي الحسني.الجزائري
- صحّح لغتك، د. ناصر لوحيشي

التقدّم

- بناء قاموس الكلمات الموسومة
- بناء وحدة اللغة العربة
- إنجاز 300 قاعدة كشف الأخطاء
- بناء مشروع التحويل الآلي للقواعد إلى صيغة XML
- دعم العربية رسميا في الإصدار رقم 4.9

مستقبلا

- إعداد قواعد أكثر.
- تحسين قاموس الوسم
- تحسين برنامج الوسم

فريق العمل

- د. صهيب عفيفي، جامعة أرتوا، فرنسا
- فلاك سمية و سعيدي ياسمين، جامعة البويرة
- كالي إيمان و تشوكشت كريمة، جامعة المدية