# arTenTen: Arabic Corpus and Word Sketches

Tressy ArtsYonatan BelinkovLexicographerMIT, USAtressy.arts@gmail.combelinkov@mit.edu

Adam Kilgarriff Lexical Computing Ltd, UK adam@lexmasterclass.com **Nizar Habash** New York University Abu Dhabi, UAE nizar.habash@nyu.edu

Vit Suchomel Masaryk Univ., Cz., Lexical Computing Ltd, UK xsuchom2@fi.muni.cz

### Abstract

We present arTenTen, a web-crawled corpus of Arabic, gathered in 2012. arTenTen consists of 5.8billion words. A chunk of it has been lemmatized and part-of-speech (POS) tagged with the MADA tool and subsequently loaded into Sketch Engine, a leading corpus query tool, where it is open for all to use. We have also created 'word sketches': one-page, automatic, corpus-derived summaries of a word's grammatical and collocational behavior. We use examples to demonstrate what the corpus can show us regarding Arabic words and phrases and how this can support lexicography and inform linguistic research.

The article also presents the 'sketch grammar' (the basis for the word sketches) in detail, describes the process of building and processing the corpus, and considers the role of the corpus in additional research on Arabic.

# 1 Introduction

Without data, nothing. Corpora are critical resources for many types of language research, particularly at the grammatical and lexical levels. In this article, we present arTenTen, a web-crawled corpus of Arabic, gathered in 2012, and a member of the TenTen Corpus Family (Jakubíček *et al.* 2013). arTenTen comprises 5.8-billion words. Since 2003, the key resource for Arabic has been Arabic Gigaword.<sup>1</sup> It contains exclusively newswire text. arTenTen improves on Gigaword, for dictionary-editing and related purposes, by covering many more types of text. A 115-million word chunk has been tokenized, lemmatized and part-of-speech tagged with the leading Arabic processing toolset, MADA (Habash and Rambow 2005; Habash *et al.* 2009), and installed in the Sketch Engine (Kilgarriff *et al.* 2004), a leading corpus query tool, where it is available for all to investigate.<sup>2</sup> There have been other important efforts in creating large collections of Modern Standard Arabic text, such as the Corpus of Contemporary Arabic (al-Sulaiti and Atwell 2006), International Corpus of Arabic (Alansary *et al.* 2007) and the Leipzig University Arabic collection (Eckart *et al.* 2014). Zaghouani (2014) has also presented a survey of several freely available corpora. These various corpora come in a range of sizes, but all of them are smaller than arTenTen.

One feature of interest in the Sketch Engine is the 'word sketch', a one-page, automatically derived summary of a word's grammatical and collocational behavior. Word sketches have been in use for English lexicography since 1999 (Rundell and Kilgarriff 2002) and are now available for twenty languages. In section 2, we describe how word sketches (and two related reports; thesaurus and 'sketch diff') can be used to give a better understanding of the behavior of Arabic words and phrases.<sup>3</sup>

To provide word sketches, we must parse the corpus either with an external parser or with the Sketch

<sup>1</sup> Arabic Gigaword is created and distributed by the Linguistic Data Consortium (Graff 2003). It is regularly updated and is now in its fifth edition.

<sup>2</sup> http://www.sketchengine.co.uk

<sup>3</sup> The methods and approach described here are similar to those used in the creation of the Oxford Arabic Dictionary (Arts *et al.* 2014)

Engine's built-in shallow parser, as here. For this process, we need a 'sketch grammar' for Arabic, which is presented in a tutorial-style introduction in section 3. Section 4 describes how arTenTen was created and prepared for the Sketch Engine. In section 5, we conclude with a summary and a brief discussion of future work.

# 2 Using arTenTen in the Sketch Engine for language research

The Sketch Engine is in use for lexicography at four of the five UK dictionary publishers (Oxford University Press, Cambridge University Press, Collins, and Macmillan), at national institutes for Bulgarian, Czech, Dutch,<sup>4</sup> Estonian, Irish,<sup>5</sup> and Slovak, and for a range of teaching and research purposes at over 200 universities worldwide.

Before discussing the details of how we built the arTenTen corpus and annotated it, we provide several examples of its utility in the context of language research, e.g., for lexicography. This section is organized around the different functions available to the linguist using the Sketch Engine to study Arabic words in their context.

# 2.1 The Simple Concordance Query function

A Simple concordance query shows the word as it is used in different texts in the corpus. Figure 1 shows the query box, while Figure 2 shows its output. A simple search query for a word such as did (child) searches for the lemma as well as the string; so, the strings الطفل (the+child), الطفل (child+their), كالأطفال (like+the+children), etc., are all retrieved.

Simple query:	طفل	Make Concordance
	Query types Context Text types	

Figure 1: Simple concordance query

<sup>4</sup> Dutch is an official language in both the Netherlands and Belgium (where it is also called Flemish), and the institute in question (INL) is a joint one from both countries.

<sup>5</sup> Much of the development work for the Sketch Engine was undertaken under a contract from Foras na Gaeilge (the official body for the Irish language) in preparation for the creation of a new English-Irish dictionary (http://www.focloir.ie). Irish is spoken in both the Irish Republic and Northern Ireland (which is part of the UK), and Foras na Gaeilge is a joint institute of both countries.

### Query طفل 71,119 (542.2 per million)

Page 1	of 3,556 GO Next   Last
http://www	كاملة ولا يستطيعان أن يكملا الشهر وهما الأن في انتظار <b>طفلهما</b> ? الأول ولا يعرفان كيف سيواجهان المصاريف الإضافية
http://www	سيواجهان المصاريف الإضافية ? أم آلام المعيلة الوحيدة <b>لطفلتها</b> التي عملت في سوير ماركت ست ساعات يوميا تقاضت عنها
http://www	شيكل في الشهر ¸ وكان عليها أن تدفع نصف معاشها لحضانة <b>لطفلتها</b> فآثرت ألا تعمل ? أم دا ف يد الذي انهار زواجه بسبب
http://adh	س ( 1 ) كيف نتواصل نحن مع  التعامل مع التوحديين الطفل التوحدي ? وكيف نساعده لكي يتواصل معنا ? ج : كي نتواصل
http://adh	التوحدي ? وكيف نساعده لكي يتواصل معنا ? ج : كي نتواصل مع ا <b>لطفل</b> التوحدي نقوم بعمل الآتي : 1 . محاولة جذب انتباه الطفل
http://adh	الطفل التوحدي نقوم بعمل الآتي : 1 . محاولة جذب انتباه <b>الطفل</b> بأسلوب واضح . 2 . استخدام وسائل وألعاب تتناسب مع
http://adh	واضح . 2 . استخدام وسائل وألعاب تتناسب مع مستوى فهم الطفل استخدام جمل قصيرة وذات محتوى بسيط من الكلمات . 3 .
http://adh	وذات محتوى بسيط من الكلمات . 4 . استخدام كلمات مستحبة المطفل ذ وتوجد عدة طرق لمساعدة >/p>. استخدام الإشارات . 5 .
http://adh	ذ وتوجد عدة طرق لمساعدة  . استخدام الإشارات . الطفل وتشجيعه في تواصله معنا وتنمية ما يبديه من تصرف سوى
http://adh	يبديه من تصرف سوى : 1 . استجابة الأم والأب إلى شد ا <b>لطفل</b> لهما نحو ما يريد . 2 . أن نكرر ما نقوله له وإعطاؤه
http://adh	نقوله له وإعطاؤه فرصة لتفهمه . 3 . تقبل وتحمل ما يقوله ا <b>لطفل</b> . حتى وان بدا ما يقوله غريبا علينا الخ
http://adh	س ( 2 ) ما هى الأمور التي تؤدي  . علينا الخ بالطفل ألتوحدي إلى التصرف السيئ أو السلوك غير المناسب كأن
http://adh	س ( 3 ) كيف نتصرف تجاه  . تغير الوجبة الغذائية الطفل التوحدي لنخبره ماذا يفعل ? وماذا نفعل عندما يقوم
http://adh	يمكنه القيام بها ? ج : من الأمور الإيجابية أن نقول المطفل ماذا يفعل , وليس ما لا يفعل . فمثلا إذا رمي الطفل
http://adh	للطفل ماذا يفعل , وليس ما لا يفعل . فمثلا إذا رمى الطفل الطعام الذي لا يريده , فعلينا أن نوضح له بهدوء أن
http://adh	لم يكن راغبا في الطعام أو يقول ( لا ) . أما إذا قام الطفل التوحدي بعمل جيد فعلينا أن نخبره أن عمله جيد ولاقي
http://adh	س ( 4 ) ما هي السلوكيات الإيجابية والمفيدة في علاج ا <b>لطفل</b> التوحدي ? وهل من الضروري وضع خطط مسبقة لكي يجيد ما
http://adh	يوجد العديد من السلوكيات الإيجابية والتي تفيد في علاج الطفل التوحدي مثل : 0 الابتسامة في وجهه . 0 الهدوء في التعامل
http://adh	وذلك له دور إيجابي في تحسن حالته فمثلا : 1 . لا يترك ا <b>لطفل</b> لاختيار ما يقوم به . 2 . اختيار الأنشطة التي يقوم
http://adh	. حتى يسهل إتمامه والنجاح فيه . ومن أمثلة ذلك : 1 ال <b>طفل</b> الذي لا يحب الازدحام يؤخذ إلى حديقة عامة قليلة الازدحام
Page 1	of 3,556 Go Next   Last

Figure 2: The resulting concordance lines

# 2.2 The Frequency functions

The Sketch Engine interface provides easy access to tools for visualizing different aspects of the word frequency (see Figures 3 and 4). The Frequency Node<sup>6</sup> forms function on the left hand menu (Figure 3) shows which of the returned forms are most frequent.

(?)	word Freq
	<u>p/n</u> الأطفال 17056
Save	<u>p/n</u> الطفل 15325
< Concordance	<u>p/n</u> طفل 4557
Sample	<u>p/n</u> أطفال 4097
Filter	<u>p/n</u> الاطفال 3465
	<u>p/n</u> للأطفال 2840
Frequency	<u>p/n</u> طفلا 1705 طفلا
Node tags	<u>p/n</u> والأطفال 1681
Node forms	<u>p/n</u> الطفلة 1587
Doc IDs	<u>p/n</u> الطفل 1518
Text Types	<u>p/n</u> طفلة 1275
Collocations	<u>p/n</u> اطفال 1155
ConcDesc	<u>p/n</u> طفاك 1022
Visualize	<u>p/n</u> والطفل <u>p/n</u>
(?)	<u>p/n</u> أطفالنا 621
	<u>p/n</u> طفلها 547 🗖
	<u>p/n</u> أطفالهم 526
	<u>p/n</u> بالأطفال 474
	<u>p/n</u> وأطفال <u>463</u>
	<u>p/n</u> أطفالها 451 ■
	<u>p/n</u> والاطفال 443 ■
	<u>p/n</u> للاطفال 443
	<u>p/n</u> لأطفال 423 ∎
	<u>p/n</u> أطفال 390
	<u>p/n</u> أطفاله 325 ∎
Menu position	<u>p/n</u> بالطفل 313 ∎
	<u>p/n</u> لطنل 282
	<u>p/n</u> طفله 260
	<u>p/n</u> فالطفل 251
	<u>p/n</u> طفاين 242

Figure 3: Frequency of node forms of طفل

The  $\mathbf{p}/\mathbf{n}$  links are for positive and negative examples. Clicking on  $\mathbf{p}$  gives a concordance for the word form, while clicking on  $\mathbf{n}$  gives the whole concordance *except* for the word form.

The Frequency Text Types function shows which top-level domain is most frequent (Figure 4).

	<u>Top level domain</u>	Freq	<u>Rel [%]</u>	
<u>p/n</u> com		38068	95.1	
<u>p/n</u> net		14110	103.8 💻	
<u>p/n</u> org		10128	118.0 💻	
<u>p/n</u> ps		947	130.5 —	
<u>p/n</u> sa		901	160.7 —	
<u>p/n</u> info		744	62.1 -	
<u>p/n</u> sy		435	126.5 —	
<u>p/n</u> ae		357	138.6 —	
<u>p/n</u> ws		338	54.4 -	
<u>p</u> / <u>n</u> edu		305	612.1	
<u>p/n</u> uk		284	87.4 —	
<u>p</u> ∕ <u>n</u> jo		271	103.8 —	
<u>p/n</u> ma		267	132.0 —	
<u>p/n</u> eg		256	71.8 —	
<u>p/n</u> sd		208	75.5 —	

المفل Figure 4: Frequency list of domain extensions of sites that contain forms of

Both hit counts and normalized figures are presented to account for the different quantities of material from different domains. If the word was equally frequent (per million words) in all of the domains, the figures in the fourth column would all be 100%. The bars are based on the normalized figures (with the height of the bar corresponding to the quantity of data). We see that علفل is frequent on *.edu* sites.

This utility is useful when researching regional differences. For example, making a frequency list for حَوْصَصة (privatization), we see (Figure 5) that it is used almost exclusively in Moroccan and Algerian newspapers.

<u>Second level domain</u>	<u>Freq</u>	<u>Rel [%]</u>	
<u>p/n</u> sawt-alahrar.net	5	516.7	
<u>p/n</u> assif.info	4	1014.8	
<u>p</u> / <u>n</u> annahjaddimocrati.org	2	18471.3	
<u>p</u> / <u>n</u> wordpress.com	1	12.0	1
<u>p</u> / <u>n</u> voltairenet.org	1	133.4	•
<u>p</u> / <u>n</u> riftoday.com	1	1759.3	
<u>p</u> / <u>n</u> odabasham.net	1	136.9	•
<u>p</u> / <u>n</u> marxy.com	1	1279.2	
<u>p</u> / <u>n</u> kassioun.org	1	191.0	•
<u>p/n</u> justgoo.com	1	544.6	-
<u>p/n</u> essaha.info	1	1469.4	
<u>p</u> / <u>n</u> educpress.com	1	834.9	-
<pre>p/n echoroukonline.com</pre>	1	867.2	-
<u>p/n</u> djazairess.com	1	98.0	1

Figure 5: Frequency list of sites containing forms of خوصصة

# 2.3 The Word List function

The Word List function allows the user to make frequency lists of many varieties. Figures 6, 7 and 8 show the tops of frequency lists for word forms, lemmas and diacritized<sup>7</sup> lemmas for the corpus.

Word list	Word list	Word list
Corpus: arTenTen12 [sample 115M]	Corpus: arTenTen12 [sample 115M]	Corpus: arTenTen12 [sample 115M]
Page 1 Go <u>Next &gt;</u>	Page 1 Go <u>Next &gt;</u>	Page 1 Go <u>Next &gt;</u>
word Freq	lemma Freq	<u>lemma_voc</u> <u>Freq</u>
<u>3242280</u> في	<u>غي 3962066</u>	<u>غي 3962066</u>
<u>2914934</u> من	<u>من 3500214</u>	مِنْ <u>3413373</u>
<u>1593477</u> على	<u>2285678</u> على	<u>2283548</u>
<u>1184760</u> أن	<u>2184612</u> أن	أن <u>1332439</u>
<u>754664</u> إلى	<u>1310294</u> الذي	<u>1310294</u> الَّذِي
<u>738288</u> عن	<u>1245137</u>	1245137
<mark>۲ <u>659851</u></mark>	<u>الى 1231294</u>	<u>1231294</u> إلى
<u>ه 637527</u>	<u>1102480</u> کان	<u>1102480</u> کان
629086 الله	⊾ <u>1009041</u>	<b>⊾</b> <u>1009041</u>
<u>610949</u> ما	۲ <u>984894</u>	<u>۶ 984894</u>
<u>585503</u> المنتي	<u>927882</u> عن	<u>927455</u> غن
<u>هذا 518842</u>	<u>899877</u> ان	<u>850948</u> أَنَّ
<u>453099</u> أو	<u>746755</u> قال	<u>746755</u> قال
<u>416753</u> الذي	<u>722219</u> الله	<u>722219</u> الله
<u>413353</u> ان	<u>640003</u>	فلك فلك
<u>402313</u> مع	ی <u>638513</u>	<u>638513</u> ق
هذه <u>402083</u>	<u>ర- 588106</u>	<u>588106</u> -لِ
<u>361499</u> کان	<u>545702</u> أو	557135 إن

Figures 6, 7, and 8: Frequency list of the whole corpus for word forms, lemmas and diacritized lemmas

### 2.4 The Word Sketch and Collocation Concordance functions

The Word Sketch function is invaluable for finding collocations. The word sketch for أخضر (green, Figure 9) shows expected collocates such as وأصفر (and yellow) and لأخضر واليابس (literally "the green and the dry"). Clicking on the number after the collocate gives a concordance of the combination (Figure 10).

<sup>7</sup> Diacritics and diacritization are often referred to as vowels and vocalization because the most common use of Arabic diacritics is to indicate short vowels. We use the more general term here to account for non-vowel diacritical marks, such as the consonant gemination marker, the shadda.

	أن					Page 1 of 9 Go <u>Next   Last</u>
ضر		arTe	enTen12 [samp	ole 115/	M] freq	http://www في كالبنانية ذ 5 كانون الأول : حرائق هائلة تقضي على الأخضر واليابس في كل لبنان والمحكمة الدولية تنفي تقريرا
and/or	<u>472</u>	0.7	adjective-of	<u>4865</u>	4.6	العولمة , الإمبراطور الأخير للحداثة المسيحية , ستأتي على الأخضر و <i>اليابس</i> في تربتنا وفي تربتهم , عدا في حالة واحدة http://isl
يابس	<u>174</u>	12.6	ضوء	<u>456</u>	10.48	والمجتمع والجماعات والأفراد في تصاعد وتيرته حتى أكل الأخضر <i>واليابس</i> ? ولا شك أن الظاهرة الاحتجاجية تدفعنا إلى http://www
أصفر	<u>52</u>	9.52	شاي	<u>178</u>	10.14	مجتمعية ممكنة اتقاء للهزات والعواصف التي تأتي على الأخضر <i>واليابيس</i> , كما حدث فعلا في بعض الأقطار العربية في http://www
برتقالي	<u>9</u>	8.45	لون	<u>300</u>	9.42	ان هذه القلعة هي المصدر الوحيد للرزق لكنهم قضوا على الاخضر و <i>البابس</i> فقد بيعت هذه الشركة بمبلغ 1.3 مليار
أزرق	<u>13</u>	8.31	خط	<u>355</u>	8.94	وجاهليته وحقارته إذا كان الهدف فتح حرب الطوائف لتحرق الأخضر <i>واليابس</i> . لكننا لا نعرف بالتحديد من يقف وراء العملية
بنفسجي	_	8.05	مسطح	<u>61</u>	8.54	" الوهم الطويل مدو ومرعب حتى لأعدائه , هو الذي أدمن الأ <b>خضر <i>واليابسية</i> " في مرحلة صعوده , ب</b> أبي إلا أن يطالهما وهو
وردي		7.88	منطقة	<u>637</u>	8.41	القرى والمدن , وجميعنا يعلم أن ظلم نظام بن على أتى على الأ <b>خضر و<i>اليابس</i> من بنزرت لبرج الخضراء , ولكن أن تصبح شاشا</b> تنا
أحمر		7.87	جبل	<u>119</u>	8.35	سكان عدد كبير من القرى إذ قضت عليها تماما وأتت على الأخضر هل المصائب التي  . ? <b>والبابس</b> معا في ميانمار
ر مادي ،	_	7.19	حزام	<u>57</u>	8.23	والاتفاق تترافق مع دعوات الى المواجهة والحرب وحرق الأخضر و <i>الما بيس</i> وترويع الناس ? وأى نظام أمنى رسمي يمكن أن
أبيض		6.99	عشب	<u>51</u>	8.1	حاق بفرص عمل ثمينة فارتموا في احضان مصارف قضت على الاخضر وبكثير من الامتعاض تقبل
جاف	_	6.78	مسيرة	83	7.76	ومن يتبعه من المتعودين على الولائم والبزنس بدرس وطحن الأخضر و <i>الباريس</i> الذي لا يمت بصلة لمشاريعهم بالاحتيال وتقاسم
أسود		6.36	مساح نبات	<u>67</u>	7.65 7.25	اليوم فلا حيلة لى سوى الصمت فبواخر هم اصبحت ناتي على الا <b>خضر <i>والما بيس</i> و</b> اناً رغم ماقدمت لازلت انا انا لا مرافق ولا
ممتاز نظري	-	5.72 5.13	علف	<u>38</u> 26	7.25	سبعة أعوام عجاف أكل المستصر وعملائه  المظلوم الأخضر و <i>اليا بيس</i> في عراقنا الجريح ودمر جميع مؤسسات الدولة
لطري مفتوح		4.48	شجرة	48	7.24	حقيقي لها , بل هي مسمى فقط , وإن هي الا فوضي أنت على الأخضر  . واليا بيس أضاعت وحدة البلاد واستقرارها وحريتها
ستوح	2	4.40	فلفل	22	7.12	القوى المختلفة , بل قد تشهد مصر حربا أهلية تأتى على الأخضر <i>والبابيس</i> , وفي ظل هذه المعطيات فإن من يدعو إلى إسقاط
			بصل	21	6.95	وان نواد هذه الفتن في مهدها قبل ان تنقشى وتقضي على الاخضر <b>واليابس</b> وتكون العواقب وخيمة و يكون قد فات الوقت
			جزيرة	57	6.94	استفاقت مدينة ليون الفرنسية على حريق أتى على /// الأخضر و <i>الباريس</i> في مستودع كبير للحافلات مملوك لشركة كيوليس
			رقعة	22	6.87	البدري ولعبه بطريقه لعب لاتناسب الفريق كانت ان تاتي على الاخضر والما يبس في الفريق وتسببت بالفعل في خروج الفريق من http://zef
			وادى	21	6.86	واحتدمت الأمور والفجرت الحرب الأهلية التي أنت على الأخضر بغادر أعضاء الفريق البرتقالي مدينة
			قبة	23	6.83	
			راية	25	6.8	Page 1 of 9 GO <u>Next</u>   <u>Last</u>
			ورق	54	6.7	
						Figure 9. Word Sketch results for أخضر (left)

Figure 9: Word Sketch results for أخضر (left) Figure 10: Concordance lines for أخضر in combination with its collocate (right)

In this concordance, we see that this combination usually occurs with 10) أتى على of the 20 lines in Figure 10) or verbs denoting destruction, such as قضى على (to destroy) for lines 1, 5, 11, and 17; and حرق (to burn) for line 10. Therefore, looking at the context, we can deduce the meaning "everything" for line idiom لأخضر واليابس the idiom أتى على لأخضر واليابس (to destroy everything). Additionally, in the Word Sketch, we see that a top collocate noun for the adjective ضوء أخضر (light). Green

Additionally, in the Word Sketch, we see that a top collocate noun for the adjective نوء is أخضر (light). Green light is not such a common phenomenon that it would account for this, so again, we look at the concordance (Figure 11).

سئول عسكري : الحكومة الإسرائيلية أعطت <b>//ضوء</b> الأخضر كثف مسئول > ع؛ writer أ ل
الكوميكس الشهيرة جدا , وهو فيلم الحركة البطولي <b>//ضوء الأخض</b> ر المنتظر عرضه خلال شهر يونيو القادم , Lantern
ور قصة فيلم <b>//ضوء</b> . بروس الأميركيتين في مصر الأخضر حول هال جوردان , وهو طيار في القوات الجوية الأمير
البريطاني ان " جمعية خيرية بريطانية حصلت على <b>//ضوء الاخض</b> ر لاطلاق لعبة يانصيب و الجائزة عبارة عن علاجات للذ
وهناك اقول كثير مكتوبة في دينهم يشجعهم ويعطيهم <b>//نصوء الاخض</b> ر في التعامل مع الاخر بكل وحشية ودموية , وهذا باطبع
اغتيال قيادات الحركة الشعبية و لكن بعد ما وجد <b>//ضوء الأخض</b> ر من الرئيس عسر البشير . في الاجتماع الذي تم بين الدك
أمريكية , ومحمية بفيتو أمريكي يعطيها أنى شاءت <i>الضوء</i> الأخضر لتواصل جرائمها على مرأى من العالم ومسمع , فهي لا
ينتظر أن يوافق مسؤولو " العميد " على إعطائه <b>//ضوء الأخض</b> ر لمسح الديون من عائدات الفريق من حقوق البث التلفزي
تومي , في الدقيقة الأولى من اليوم 5 جويلية , <b>//صوء الأخض</b> ر لانطلاق فعاليات المهرجان الثقافي الإفريقي , بقصر ,
الاثنين الفارط وأشعره بقراره الأخير القاضمي بإعطاء <b>//ضوء الأخض</b> ر للتشكيلة البليدية للعودة إلى ملعب تشاكر أوراق المدرب
نفت مصادر سياسية رفيعة المستوى ان يكون هناك اي <b>ضوء</b> ا <b>خض</b> ر  . يتعلق برئاسة مجلس النواب , وتفضيل مرشح
من لبنان وتوطين الفلسطينيين في لبنان وإعطاء <b>//ضوء الأخض</b> ر لإسرائيل بضم الضفة الغربية رسميا , أو عمليا على الأ
جان يفلت من العقاب نتيجة لتخاذل السلطات يشكل <b>ضوء/ أخضر</b> بأن هذه السلطات لن تبالي بمحنة ضحايا العنف الجنسي
ح كيفن راد <b>/لصوء</b>
اللجنة العليا الصعود لتنفيذ التمرين بدون إشارة أو <b>ضوء</b> أ <b>خض</b> الدرجة النهائية = صفر مخالفات الفريق تنافس الجمبازي
أصحابها عن أن السيد رئيس بلدية بنكرير قد أعطى <b>//ضوء الأخض</b> ر للمواطنين بالبناء والإصلاح دونما الحصول على رخصه
سلاحه إلى أخيه سواء كان عنصرا أو قياديا يعطي <i>الضوء</i> الأخضر للقتل ( لان كلاهما قاتل ) حتى يسفك الدم الذي تباكر
ن رب العمل قد أعطاني <b>//ضوء</b> السؤال > الأخضر لأخذ ما يكفيني من أرباح أمواله الذي هو خاصته , لكن
بخفة , ينادي لبيع بضاعته , وحين تضيء الإشارة <i>بالضوء</i> الأخضر يهم بالابتعاد خوفا من دهس بغير حساب , وكثيرا ما ,
والتلفاز والصحف . وقد أعطت إدارة بوش رايلي <b>//ضوء الأخض</b> ر لتشغيل إذاعة العراق الحر . رايلي يرتبط بخطة إدارة
ضوء in combination with أخضر Figure 11: Concordance lines for

In these lines, we can see that الضوء الأخضر (the green light) is used in much the same way as the English, in "to give/get the green light", meaning to be allowed to go forward.

# 2.5 The Bilingual Word Sketch function

A new function of the Word Sketch is the bilingual word sketch, which allows the user to see word sketches for two words side-by-side. Figure 12 shows a comparison between أحمر and *red*.

دن Some of the same things are أحصر /red in Arabic and English; thus, we find the matched pairs المحمر/meat, المحمر/carpet, and المحمر/pepper. All three are to an extent idiomatic, with the same idiomatic meaning in both languages. The *Red Cross* and *Red Crescent* are discussed more in Arabic media than in English, reflecting the unfortunate reality of several Arabic-speaking countries today. In contrast, *wine* is high in the English list but absent in the Arabic one.

به and English أحمر ed

adjective-of		4.8	modifies		0.4
صليب	<u>1132</u>	11.52	flag	<u>27161</u>	9.32
ھلال	<u>1101</u>	11.38	carpet	<u>21070</u>	9.04
بحر	<u>1498</u>	11.0	wine	<u>34956</u>	8.8
خط	<u>845</u>	9.92	tape	<u>15035</u>	8.44
لون	<u>507</u>	9.75	meat	<u>17594</u>	8.34
فلعة	<u>214</u>	9.19	pepper	<u>10687</u>	8.26
بطاقة	<u>249</u>	9.08	herring	<u>6067</u>	7.93
لحم	<u>208</u>	8.87	light	<u>25976</u>	7.58
هندي	<u>157</u>	8.6	onion	<u>5258</u>	7.33
ساقية	<u>93</u>	8.24	rose	<u>4714</u>	7.3
دم	<u>212</u>	7.95	cell	<u>16369</u>	7.22
كرية	<u>75</u>	7.93	lipstick	<u>3047</u>	6.86
سجاد	<u>77</u>	7.92	ink	<u>3605</u>	6.73
شمع	<u>67</u>	7.76	bump	<u>2858</u>	6.59
شيطان	<u>87</u>	7.7	grape	<u>2877</u>	6.51
زاوية	<u>69</u>	7.54	stripe	<u>2462</u>	6.43
ياقوت	<u>47</u>	7.23	ribbon	<u>2586</u>	6.41
فلفل	<u>46</u>	7.22	sole	<u>2300</u>	6.39
ورد	<u>58</u>	7.07	lip	<u>3592</u>	6.35
طوب	<u>34</u>	6.78	shirt	<u>4524</u>	6.33
زاءوق	<u>32</u>	6.72	berry	<u>2563</u>	6.31
كبريت	<u>32</u>	6.71	hair	<u>9863</u>	6.29
خمير	<u>32</u>	6.71	dress	<u>6136</u>	6.24
درپ	<u>34</u>	6.64	snapper	<u>1856</u>	6.24
بقعة	33	6.62	arrow	2207	6.19

# 2.6 The Distributional Thesaurus function

The Sketch Engine also offers a distributional thesaurus, where, for the input word, the words 'sharing' the most collocates are presented. Figure 13 shows the top entries in similarity to تصدير (export). The top result is (import). Clicking on this word takes us to a 'sketch diff', which is a report that shows the similarities and differences between the two words in Figure 14.

	**		اد	ستير	١/	دير	ص	۔ ڀ ڌ	rTenTen12 [sampl	e 115M	] freas	= 503	7/36	11				
ىدىر	لص	arTenTe	ىدىر	6.0 تم	4.0		.0	0	-2.0 -4.0	-6.0	أستيراد			_				
Lemma	Score	Freg	s	ubject-of	289	252	1.3	1.7	construct-state	1892	1129	2.0	1.9	and/or	602	412	1.6	1.
		· ·	ر	حظر	<u>3</u>	<u>8</u>	4.8	6.3	مشتق	0	<u>12</u>		7.5	تصدير	0	<u>261</u>		11
<u>آستير اد</u>	0.459	3611	ز	جار	<u>3</u>	<u>5</u>	3.8	4.5	خضار	0	<u>9</u>		7.1	جمرك	0	<u>3</u>		5
تسويق	0.343	5379	تم	ý.	<u>48</u>	<u>66</u>	6.4	6.9	بنزين	0	<u>8</u>		6.8	شراء	0	<u>5</u>		4
<u>تخزين</u>	0.322	2285	تم	1	<u>74</u>	<u>99</u>	4.9	5.3	خضرة	0	<u>8</u>		6.7	نداول	0	<u>3</u>		4
تصنيع	0.307	3483	ام	ک	<u>10</u>	<u>9</u>	4.2	4.1	لباس	0	Z		6.7	توزيع	<u>5</u>	<u>15</u>	4.3	5
أستهلاك	0.3	5377	ى	جر;	Z	<u>5</u>	4.5	4.0	لحم	2	<u>52</u>	5.5	8.2	أستهلاك	<u>10</u>	<u>6</u>	6.3	5
<u>توريد</u>	0.287	1190		کار	<u>10</u>	0	0.6		دواه	<u>8</u>	<u>40</u>	4.9	7.3	بيع	<u>11</u>	<u>6</u>	4.9	4
أستخراج	0.266	2456		مکر	<u>15</u>	<u>8</u>	4.0	3.1	زيت	<u>6</u>	<u>18</u>	5.3	7.1	نقل	<u>19</u>	<u>10</u>	4.6	3
شراء	0.265	15838		مذ	<u>11</u>	<u>5</u>	5.5	4.3	قمح	<u>14</u>	<u>37</u>	7.0	8.7	صناعة	2	3	4.3	
تهريب	0.254	3817		أستطا <u>.</u>	3		3.0		سلع	<u>18</u>	<u>41</u>	6.5	7.9	تسويق	<u>18</u>	4	7.1	5
<u>تجارة</u>		17570		وقف	4	0	3.3		کمیة	22	33	6.3	7.0	ا <sup>لت</sup> اج	<u>63</u>	<u>14</u>	6.7	4
<u>بر۔</u> إنتاج		28181	اد دأ	أرا	5		3.9		سلعة أسمنت	<u>8</u>	<u>8</u>	6.2 7.6	6.5	سعر	3	0	2.3	
	0.240	2534		ب توقف	<u>9</u> 3	0	4.0 4.1		اسمت	<u>20</u> 29	<u>18</u> 25	7.5	7.8 7.5	تجارة أستثمار	Z 18	0	3.7 5.3	
<u>تجميع</u>				بوھ أمكر	<u>2</u> 4	0	4.1		بطاعة أرز	15	<u>25</u> 7	7.5	7.5 6.4	استمار تحمیل	3	0	5.3 5.4	
توزيع		16139		امتر حاوا			4.9		منتوج	<u>15</u> 10	<u>/</u> 3	6.8	5.5	تحضين	2 4	0	5.4	
<u>ترويج</u>	0.237	4061		ـــر. أعا	<u>/</u> 4	0	5.7		سرج الملاح	83	26	6.8	5.1	تخزين	- - 	0	6.2	
<u>إدخال</u>	0.235	6883		ت.	3	0	6.6		منتج	84	24	7.7	6.0	سرمن شحن	9	0	6.8	
<u></u>	0.226	28032			- 6	0	7.6		نفط	219	39	8.7	6.2	توريد	4	0	6.8	
<u>تداول</u>	0.221	9957		ار۔ أحتك	3		7.6		غاز	443		10.3	7.3	تکریں	5	o	7.4	
أستبدال	0.197	3251			141	-		0.3	يترول	36	3	7.7	4.4	تصنيع	 17	0	7.7	
نوعية	0.196	4059		ajective ترف	141	129	0.2	9.0	ار هاب ار هاب	42	0	6.6		آستخراج	2	0	7.7	
تدفق	0.194	3801		ىرم عشوائم	0	<u>د</u> 13		7.4	ئورة	134	0	6.8		آستيراد	<u>261</u>	0	11.7	
تفريغ	0.189	1618		طسو،م أستهلاكم	0	5		6.9	شحنة	10	0	6.9						

Figure 13: Thesaurus search showing entries similar to تصدير (export) (left) Figure 14: Sketch Diff comparing collocates of استيراد and استيراد (export and import) (right) The first number following the collocate shows the number of occurrences of this collocate with تصدير, the second number shows the number of occurrences with استيراد. A color scale from green to red visualizes the distribution.

### 2.7 Collocations and lexicographic research: Two case studies

The information in the Sketch Engine reports is particularly useful for lexicographers. It presents collocations, idioms, prepositions commonly occurring with verbs, and so forth.

It also gives insight into the use of words, often assisting the lexicographer in finding definitions for new words, for example, for توحي (autistic), as shown in Figure 15. The immediate context of *child* and *patient* indicate that the word might be an adjective for an ailment.

وسبل, illuminative , الروح الى الله من خلال المسهل توحدي ويوجه الرسالة البابوية لاوون الثالث عشر ع <

It also occasionally reveals new senses of words. For example, the word نسق is traditionally known to mean "order/manner", as illustrated in Figure 16.

نسق nasaq order, array, layout, arrangement, disposition; connection, succession, sequence; manner, mode, system, method; symmetry; نسق nasaqan in regular order, in rows | نسق nasaqan in the manner of; على نسق واحد in the same manner, equally, evenly, uniformly; حروف النسق

النَّسَقُ - نَسَقُ : النَّسَقُ : ما كان على نظام واحد من كل شيء . يقال : جاءَ القومُ نَسَقًا ، وزُرعَت الأَشجار نَسَقًا . ويقال : شَعْرٌ نَسَقٌ : مستوي النِّبتة حَسَنُ التَّركيب ، ودُرٌّ نَسَقٌ : منتظمٌ . و النَّسَقُ المنسوق . يقال : كلامٌ نَسَقٌ : متلائم على نظام واحد . و ( حروفُ النّسَق ) : حروفُ العطف . **المعجم:** المعجم الوسيط

Figure 16: Dictionary entries for نسق from Wehr's Dictionary of Modern Written Arabic 4th ed. 1979, and *al-mu'jam al-wasit* (Academy of the Arabic Language in Cairo)<sup>8</sup>

However, looking at the concordance for the top adjective collocate تساعص (increasing, Figure 17), we see that these sentences do not seem to refer to "increasing order" but to an "increasing pace"

للتجار والقطاعات . فقد عرف حجم المشاركة من قبل التجار نسقا رضي تصاعديا حيث تجاوز السنة الفارطة الألف تاجر واتسعت قبل عشيرة وأقارب محمد البوعزيزي لتأخذ هذه المظاهرات نسقا رضي تصاعديا مع انتحار شاب آخر بصعقة كهربائية في 22 ديسمبر الشغالات ( النحل ة العاملة ) يبدأ بالتناقص تدريجيا وينسق رضي تصاعدي وسريع ينتهي بضعف ملفت للخلية ثم إتلافها نهائيا الكبير في تتوعه . وكذلك الإستعمال المكثف والعشوائي وينسق رضي تصاعدي للمبيدات الحشرية الفتاكة في النشاط الفلاحي فرنسا وألمانيا وكندا بمكاتب تسجيل متنقلة مشيرا إلى النسق رضي التصاعدي الماحوظ في عمليات تسجيل التونسيين بالخارج كاكا " أحلى الأوقات " مع ريال مدريد هذه الفترة , بعد النسق رضي التصاعدي الذي طرأ على أدائه منذ بداية الموسم الجاري Figure 17: Concordance for نسق with

<sup>8</sup> Entry as found at <u>almaany.com</u>, February 2014.

Investigating the word further, we find that "pace" is a common contemporary meaning of the word.

Having shown the functions of the Sketch Engine and its functionality for Arabic, we will now go into more detail on developing the corpus and deploying it in the Sketch Engine.

# 3 A Sketch Grammar for Arabic

A sketch grammar is a grammar for the language based on regular expressions over part-of-speech tags (see Kilgarriff *et al.* 2004). It underlies the word sketches and is written in the corpus query language (CQL). A sketch grammar is designed particularly to identify head-and-dependent pairs of words (e. g., تستق, تساعدي ) in specified grammatical relations (here, adjective-modifier) so that the dependent can be entered into the head's word sketch and vice versa. Prior to the work described here, there has only been one sketch grammar for Arabic, developed at Oxford University Press (OUP) as part of the development phase for the Oxford Arabic Dictionary (Arts *et al.* 2014). It (and the word sketches resulting from it) is accessible only on arrangement with OUP.

The sketch grammar is one of the two components needed to build word sketches. The grammar is run over the corpus to identify all of the <word1, grammatical-relation, word2> triples in the corpus. The other component is a statistic. For each lemma occurring in the word1 slot (the node word) and for each grammatical relation, we count the number of times each different lemma occurs in the word2, or 'collocate', slot. We use these numbers to calculate an association score<sup>9</sup> between the node word and the collocate. The collocates with the highest association scores go into the word sketch.

A sketch grammar contains a set of definitions for grammatical relations. A simple grammatical relation definition is just:

# =adjective 1:"noun" 2:"adj"

This definition says that if we have a word with part-of-speech tag *noun* followed by one with part-of-speech tag *adj*, the grammatical relation *adjective* holds between the node word (the noun) and the collocate (the adjective). The 1: identifies the noun as the first argument of the grammatical relation, and the 2: identifies the adjective as the second argument.

We would also like to identify the noun as a collocate, when the adjective is the node word. To do that, we tell the system that the relation is *dual* and give a name for the inverse relation: here, *adjective-of*, as follows.

\*DUAL =adjective/adjective-of 1:"noun" 2:"adj"

There is some shorthand here. There may be many different fields of information associated with a word, of which the part-of-speech tag is just one field. In the case of arTenTen, there are many fields, including the word form itself, the lemma (with and without diacritics), the case and the state.<sup>10</sup> The part-of-speech tag is called simply *tag* and in the formulation above, this has been set as the default. A non-shorthand version is

```
*DUAL
=adjective/adjective-of
1:[tag="noun"] 2:[tag="adj"]
```

All of the constraints on a word (or, technically, a *token*: tokens are usually either words or punctuation) are placed within square brackets, and each square-bracketed item relates to one token in a sequence.

10 See also section 4.2.

<sup>9</sup> The association score currently in use is a variant of the Dice coefficient; see Rychlý (2008) for full details.

Now, the linguist will immediately note that there are many cases where adjectives happen to follow nouns but are not their modifiers. The definition above is insufficiently constrained and will give rise to many false positives. One constraint we want to add is that the adjective and noun agree, in case and in state. This is enforced in the next version.

\*DUAL
=adjective/adjective-of
1:[tag="noun"] 2:[tag="adj"] & 1.state = 2.state & 1.case = 2.case

Now, an adjective followed by a noun only matches if the *state* value of the token indexed by 1: is the same as the *state* value of the token indexed by 2:, and likewise for *case*.<sup>11</sup>

This is better and will not include many false positives. However, we should also be alert to valid cases of adjectives modifying nouns, which the definition above misses. One case is where two adjectives in succession modify a noun, e.g., السعودية المركة العربية (lit: the Saudi-Arabian Kingdom). Only the adjective closest to the noun is captured by the clause above. To capture the other adjective, we add another clause to the definition:

```
1: [tag="noun"] [tag="adj"] 2:[tag="adj" & pref1tag!="prep"] &
1.state = 2.state & 1.case = 2.case
```

This version allows an intervening adjective between the noun and its collocate adjective, which must not have a prefixed preposition.

The process of developing a sketch grammar is supported by the Sketch Engine because the CQL queries can be posed directly to the corpus, using the 'CQL' option in the concordance form. Thus, the strings above can be cut and pasted into the CQL box (Figure 18), and the developer can immediately see all of the hits (Figure 19).

Simple query:		Make Concordance	
	Query types Context Text types		
Query type	$\odot$ simple $\bigcirc$ lemma $\bigcirc$ phrase $\bigcirc$ word $\bigcirc$ charac	ter 🖲 CQL	
Lemma:			
Phrase:			
Word Form:		match case	
Character:			
CQL:	1:"noun" "adj" 2:"adj" & 1.state = 2.state & 1.cas	Default attribute: tag	▼ <u>Tagset summary</u>
Make Concor	dance Clear All		

Figure 18: Using CQL in the concordance search form (with *tag* as default attribute)

<sup>11</sup> Gender and number may seem to be good candidate features for this sketch grammar. However, since MADA uses what Habash (2010) terms *form-based* gender and number, and given the prevalence of deflected agreement (irrational plural nouns take feminine singular adjectives), these features are not good indicators of noun-adjective agreement. For more on issues of Arabic agreement, see Alkuhlani and Habash (2011).

Page       1       of 2       Go       Next   Last         307351       محد افعال حقيقية من قبل واشنطن       محد افعال حقيقية من قبل واشنطن         417451       .       .         544001       9       .         .       .       .	وواعنطن ستنفذ كل الخطوات الضرورية لوقف المشروع الذري الإيراني فاذا لم تو تسمع عن الفساد والفوضى والحرب فى المسعده والاختلالات السياسية والامنية فى البلاد وتطويرها تحت شعار ( نحو منهج تربوي متطور لبناء الانسان العراقي الجديد ونكرت ه
جد افعال حقيقية من قبل واتسنطن . 417451 استيقظ من سباتك !! سنة و 9 .	وواشنطن ستنفذ كل الخطوات الضرورية لوقف ال <b>مشروع الذري الإيراني</b> فاذا لم تو تسمع عن الفساد والفوضى والحرب فى المسعده والا <b>ختلالات السياسية والامنية</b> فى البلاد وتطويرها تحت شعار ( نحو منهج تربوي متطور لبناء الا <b>لاسان العراقي الجديد</b> ونكرت ه
استَيقظ من سباتك !! سنة و 9	تسمع عن الفساد والفوضى والحرب في الصعده والافتلالات السياسية والامنية في البلاد وتطويرها تحت شعار ( نحو منهج تربوي متطور لبناء الالسان العراقي الجديد وذكرت ه
	وتطويرها تحت سعار ( نحو منهج تربوي متطور لبناء الاسان العراقي الجديد وذكرت ه
accepted in the second off in the second	
مصادر صحفية ان وزير التربية  . ( 666351	the second s
على اننا امة 862751 ", 23/12/2008	", عبدالرحمن محمد /p>= !! وهذا واحد منهم المملكة العربية السعودية هذا دليل
نفيذ هجمات على اجهزة الامن في 1121401	رجال قالوا انهم تلقوا أموالا وأسلحة من <b>ئانب لبناني مثاهض</b> لسوريا ل <sup>ت</sup>
لحرة , والتي أبرم بعضبها في عام 2150051	قسرا بما تقول فما الحال إذن في عقود منح المصائع المحلية الثابعة المنطقة ا
التجريح مو لهدف ان ما نزعل أحد 2821501	هالفكرة عفا عليها الزمن علينا مراجعتها مراجعة دقيقة وصريحة مو لهدف
ت هذا الاجتماع , ان هذا الاجتماع	لرصد العلاقة بين ارهاصات التفكير بحداثة للا <b>جتماع العربي الاسلامي</b> وبين تراد
الأغنية العالمي . ولخطورة وجود هذا 3559851	تحتوي على نسبة تغوق الحد المسموح به من قبل <b>المواصفات السعودية والأمريكية</b> ودستور ا
صياع وراء نصيحة قريب بدت عليه , 4803851	فكم من عارض صحى يبدو تافها يكون مؤسّرا , لأزمة صحية حقيقية وعدم الإن
ها هذه المرة ريتشارد قلب الأسد (1846301	الموجهة إلى العراق في سُهر رمضان على أنها <b>درب صليبية جديدة ,</b> لا يقود
تمد فيها على مقدمة قانون التوزن 5737151	- قانون الاعمار وهو قانون 10 >. الانتخابات برؤية علمية وعصرية جديدة يعة
ح أصدقائها اللبنانيين ستدفع في , 6352651	نبيه بري اعتقاده أن " الأمور ستتبلور خلال الأيام القليلة المقبلة وفرنسا م
ت التي يفترضيها التأريخ أو النقد 6510601	فائزين في هذا العمل , ولا أي تسكل من أسكال ا <b>لبحث المنهجي الملائم</b> للتصنيفان
/p>2011 2011 < ورو على الموظفين الحيتان 8194451	ليسّ كل الهجوم على معاليه مع انه للامانه السان متواضع ونظيف 10-11-
هذه القضية هو أن المعتدي غالبا 8244451	وأوضح أن

Figure 19: Resulting concordance with noun-adj-adj sequences

Typically, this will include false positives, and the developer can then add constraints to rule them out. They should also think about the cases they are missing (in this example, the two-adjective case) and need to aim for as large a population of hits as possible, without too many false positives. In the terminology of information theory, they need to attend to recall - missing items that should be found - as well as precision - avoiding false positives. Recall tends to be a harder problem because a tool cannot show the items that are not found.

The Arabic sketch grammar aims at identifying the main grammatical relations while ensuring high-quality results. The grammatical patterns it covers are:

• **subject**, **subject**-of: These relations capture the relationship between verbs and their subjects. The noun is required to appear in the nominative case and may not have a prefixed preposition or conjunction.

The phrase نزل المطر (fell) produces two grammatical relations. When نزل المطر (fell) is the node word, the grammatical relation *subject* holds between it and its collocate المطر (rain). Conversely, if المطر is the node word, then it stands in the grammatical relation *subject-of* with المطر.

• **adjective**, **adjective**-of: These two relations capture noun-adjective pairs. We enforce agreement in state (definite/indefinite) and case. Enforcing agreement in gender and number is not trivial and left for future versions.

In the phrase بحث علمي (scientific research), the noun علمي takes the *adjective* علمي, which itself is *adjective-of* for ...

• **construct-state**: Captures construct state (idafa) constructions between two nouns. The first noun is required to be in the construct state and the second noun is required to be in the genitive case with no prefixed preposition or conjunction.

In the phrase مدير المدرسة (the school principal), the grammatical relation *construct-state* holds between the node word مدير (principal) and the collocate المدرسة (the school).

• **and/or**: This relation captures conjunctive constructions of pairs of nouns, adjectives, and verbs. We enforce agreement in certain grammatical features between the two words: for nouns and adjectives, we enforce agreement in case and state; for verbs. in aspect. This relation is declared as *symmetric*, which tells the system that both words can be the head node in turn.

Examples for pairs of adjectives include: كبير وصغير (large and small) and كبير أو صغير (large or small). In these examples, the word صغير (large) stands in grammatical relation of *and/or* with صغير (small) and vice versa. Similarly, we obtain pairs of nouns (e.g., النساء والرجال, "women and men") and verbs (e.g., يضحك أو يبكى

The grammar focuses on the highest-confidence patterns for each grammatical relation. There are many constructions it does not yet cover. The quality of the identification of the different relations depends on the correctness of the automatic disambiguation component. Since the accuracy of automatic prediction of case is somewhere in the mid 80%, we can expect a fair amount of failed matches, e.g., verb-object pairs analyzed as verb-subject pairs. Future versions will increase coverage for current relations and add additional relations such as **verb-preposition** and **direct-object**. See Appendix 1 for the full grammar and the Sketch Engine documentation<sup>12</sup> for a full account of the formalism.

# 4 Creating and preparing the corpus

# 4.1 Crawling and text preparation

The following describes the processing chain for creating the corpus.

- We use texts from Arabic Wikipedia and other Arabic web pages to build the language-specific models that we need: (a) a character trigram model for language identification, (b) a byte trigram model for character encoding detection, (c) the most common Arabic words for seeding the crawl and for distinguishing sentences from lists and headers, and (d) parameters for the boilerplate cleaning utility.
- We crawl the Arabic web with SpiderLing<sup>13</sup> (Pomikalek and Suchomel 2012), a crawler designed specifically for preparing linguistic corpora. The seeds for the crawl were generated by taking the top 1000 words from Arabic Wikipedia, randomly combining them into triples, and using the triples as Yahoo queries. The Yahoo search hits gave 4583 URLs, which were used as starting points for the crawl.
- We remove the non-textual material and boilerplate with jusText (Pomikalek 2011). JusText uses the working definition that we want only 'text in sentences' (excluding e.g., headers and footers). The algorithm is linguistically informed, rejecting material that does not have a high proportion of tokens that are the grammar words of the language; therefore, in the course of data cleaning, most material, which is not in the desired language, is removed.
- We de-duplicate with Onion (Pomikalek 2011) to remove near-duplicate paragraphs. We de-duplicate at the paragraph level because for many linguistic purposes, a sentence is too small a unit, but a whole web page (which may contains large chunks of quoted material) is too large.

These tools are designed for speed and are installed on a cluster of servers. For a language where there is plenty of material available, we can gather, clean and de-duplicate a billion words a day. ArTenTen was collected in 14 days. Table 1 presents the various statistics from arTenTen.

Data statistics	Documents (web pages; millions)	Sentences (millions)	Words (millions)	Data size
HTTP requests issued	87.8	_	_	_
Web pages received	58.8	_	_	2,015 GB
Cleaned text without exact duplicates	21.5	463	17,500	152 GB

<sup>12</sup> http://www.sketchengine.co.uk/documentation

<sup>13</sup> http://nlp.fi.muni.cz/trac/spiderling

Final text without near duplicates	11.5	177	5,790	58.0 GB
Processed with MADA	0.23	4.5	115	1.32 GB <sup>14</sup>

### 4.2 Composition

The best-represented top level web domains in the corpus are .com, .net, .org, .info, .ps (Palestine), .sa (Saudi Arabia), .sy (Syria), .eg (Egypt), and .ae (United Arab Emirates), as shown in Table 2. There are 116,000 web domains represented by at least one document, and 43,000 represented by at least 10 (see Table 3), suggesting a heterogeneous corpus in contrast to corpora such as Arabic Gigaword or KSUCCA (Alrabiah *et al.* 2013), which are built from a small number of sources. The twenty domains that contributed the most documents are given in Table 4.

TLD	%	Note
.com	54.45	Generic commercial
.net	20.86	Generic network
.org	10.32	Generic organization
.info	1.69	Generic information
.ps	1.55	Palestine
.sa	1.41	Saudi Arabia
.sy	0.76	Syria
.eg	0.61	Egypt
.ae	0.60	United Arab Emirates
.CC	0.43	Cocos Islands/generic
.uk	0.41	UK
.cn	0.41	China
.jo	0.40	Jordan
.sd	0.38	Sudan
.ma	0.35	Morocco
.lb	0.30	Lebanon
.il	0.28	Israel
.biz	0.26	Generic business
.ws	0.26	Samoa/generic
.ir	0.25	Iran
Other	4.03	

Table 2: Document (web pages) by top-level domain (TLD)

>= 1 document	116,029 websites
>= 10 documents	43,282 websites
>= 100 documents	11,242 websites
>= 1,000 documents	2264 websites
>= 10,000 documents	112 websites

Table 3: Distribution of documents by website

aawsat.com	28,689
maghress.com	24,925
masress.com	23,818
sawt-alahrar.net	22,669
burnews.com	21,474
humum.net	21,084
chelseafarms.com	20,216

14 The size of the annotated corpus is 1.32 GB without morphological tags and 23.6 GB with full MADA morphological annotation.

nabanews.net	19,490
sarayanews.com	17,534
algomhoriah.net	17,090
anhri.net	16,718
tayyarcanada.org	16,315
arabic.xinhuanet.co	
m	15,879
alsahafa.sd	15,774
m.islamweb.net	15,600
digital.ahram.org.eg	15,487
arabtimes.com	15,339
rosaonline.net	15,266
alwasatnews.com	15,210
elbiladonline.net	14,934

Table 4: Websites contributing the most documents

### 4.3 **Processing with MADA**

We chose to use the MADA tool for Arabic processing because of its state-of-the-art results on Arabic disambiguation, part-of-speech tagging and lemmatization and its holistic approach to modeling Arabic, predicting all of a word's morphological features in context. MADA has been successfully used by numerous Arabic NLP projects: in the NIST Open machine translation evaluation in 2012, nine out of twelve teams competing on Arabic-English translation used MADA. In a precursor to the work described in this article, Oxford University Press used MADA to prepare corpus materials used to create the Oxford Arabic Dictionary (Arts *et al.* 2014).

Within the framework of Arabic processing via MADA (Habash and Rambow 2005; Habash *et al.* 2009), we need to distinguish two concepts: **morphological analysis** and **morphological disambiguation**. **Morphological analysis** refers to the process that determines for a particular word all of its possible morphological analyses. The word, for MADA, is the orthographic word, defined as the sequence of letters delimited by spaces and punctuation. In Arabic, the word may include a variety of clitics, such as the definite article, prepositions, conjunctions and pronominals.

Each single analysis (out of many) includes a single choice or reading of the word with multiple dimensions of morphological information: the word's full diacritization, lemma, stem, part-of-speech (POS); the full Buckwalter Analyzer tag (Buckwalter 2002), values and POS tags for four possible proclitic slots; the values of eight inflection features -- person, aspect, voice, mood, gender, number, state and case; enclitic value and POS tag; English gloss; and whether the word had a spelling variation. Table 5 shows the MADA features for the example word ويفكرة *wbfkrp* assuming a specific analysis corresponding to the English 'and with an idea'.

MADA Feature	Explanation of Feature
diac:wabifikorapK	التشكيل Diacritization
lex:fikorap_1	المفردة Lemma
stem:fikor	الجذع Stem
pos:noun	قىم الكلام Part-of-speech
BW:wa/CONJ+bi/PREP+ fikor/NOUN+ap/NSUFF_FEM_SG+K/CASE_INDE F_GEN	قسم الكلام بنظام باكوالتر Buckwalter POS tag
prc3:null	Third proclitic position away from base word (typically, interrogative Hamza) لحَاة \سابقة استغهام
prc2:wa conj	حرف \ سابقة عطف Second proclitic position away from base word

prc1:bi_prep	حرف \ سابقة جر First proclitic position away from base word
prc0:0	Zeroth proclitic position away from base word (typically the determiner Al) ال \ سابقة التعريف
per:na	Person (not applicable here) الشخص
asp:na	Aspect (not applicable here) الزمن
vox:na	معلوم/مجهل) البناء) ( voice (not applicable here
mod:na	الصيغة (not applicable here) الصيغة
gen:f	Gender (feminine here) الجنس
num:s	Number (singular here) العدد
stt:i	التعريف (indefinite here)
cas:g	لحالة لإعرابية (genitive here)
enc0:0	only enclitic after the base word ضمير \لاحقة متصل
spvar:lex	Spelling Variant (none, exact lexicon match here) إملاء غير قيلىي
gloss:idea;notion;concept	English gloss

Table 5: MADA analysis of وبفكرة wbfkrp

Arabic words are highly ambiguous, primarily because diacritical marks are usually left out. A good analyzer produces the full set of choices for a particular word out of context. For example, the word بين *byn* can have many analyses, including:

Diacritization	Buckwalter POS tag	English Gloss
bay~an+a	PV+PVSUFF_SUBJ:3MS	He demonstrated
bay~an+~a	PV+PVSUFF_SUBJ:3FP	They demonstrated (f.p)
Biyn	NOUN_PROP	Ben
bay~in (dropping all case endings for simplicity)	ADJ	Clear
Bayn	PREP	Between, among

**Morphological disambiguation** refers to selecting the appropriate morphological analysis in context. Compare the following two sentences, which both contain بين *byn*. A good disambiguation model would select the proper noun reading for (1) and the preposition reading for (2):

(1) هل سينجح **بين** أفليك في دور باتمل؟ Will **Ben** Affleck be a good Batman?

Kerry tries again to save the negotiations **<u>between</u>** Palestine and Israel.

The task of morphological disambiguation for English is referred to as POS tagging because for English, a large part of the challenge is to determine what a noun, verb, or adjective is (for example, for base forms such as *promise*, s-forms such as *promises*, ing-forms such as *promising* and ed-forms such as *promised*.). The standard English POS tag set, although only comprising 46 tags, completely disambiguates English morphologically. In Arabic, the corresponding tag set comprises thousands of tags, so the task is considerably harder. Reduced tag sets have been proposed for Arabic in which certain morphological differences are conflated, making the morphological disambiguation task easier. The term POS tagging is usually used for Arabic with respect to some of the smaller tag sets (Habash 2010).

MADA uses a morphological analyzer for MSA based on the standard Arabic morphological analyzer (SAMA) (Graff *et al.* 2009). It also uses a set of different classifiers that classify the values

of specific features from the analysis form in context, such as lemmas or gender. These features are trained on the Penn Arabic Treebank (Maamouri *et al.* 2004). The two sets of information (out-of-context analyses and in-context classified features) are combined to select the appropriate analysis in context (Habash and Rambow 2005; Roth *et al.* 2008).

A 115-million word subset of arTenTen was processed with MADA. The single preferred analysis for each word was output and used as the input to the next process. The work on MADA has been extended to handle Arabic dialects, specifically Egyptian Arabic (Habash *et al.* 2013). However, in this work, we only use MADA for MSA.

# 4.4 Into the Sketch Engine

Loading the arTenTen into the Sketch Engine required a conversion of MADA output into the format specified by the Sketch Engine. The Sketch Engine input format, often called "vertical" or "word-per-line", is as defined at the University of Stuttgart in the 1990s and is widely used in the corpus linguistics community. Each token (e.g., word or punctuation mark) is on a separate line and where there are associated fields of information, such as lemma, POS-tag and morphological features, they are included in tab-separated fields. The conversion script extracts all of the MADA-generated features into fields and incorporates additional fields for ease of search in Sketch Engine, e.g., Arabic-script, diacritized and non-diacritized versions of the lemma (back-transliterated from the Buckwalter transliteration). Structural information, such as document beginnings and ends, sentence and paragraph mark-up, and any available metadata, are presented in XML-like form on separate lines. For web corpora, there is limited metadata available; date of collection and the URL from which the domain and top-level domain can be derived are useful. A sample of the vertical file is shown in Appendix 2.

In the Sketch Engine, each corpus has a corpus configuration file, which specifies the information fields that the corpus includes and various aspects on how they should be displayed. The next stage of the corpus preparation was to develop the arTenTen corpus configuration file. For instance, we needed to specify here that the word sketch attribute is the Arabic form of the lemma to facilitate searching by users in Arabic. This was problematic: it was not clear whether this should be the version of the lemma with diacritics or without. The no-diacritic option was desirable simply because it was the way that Arabic speakers usually write. If we did not permit no-diacritic input, beginner users would obtain no results and would be put off. However, if the diacritics are not written, the level of ambiguity is considerably higher, and it would not be possible to see a word sketch for  $\hat{J}$  (to confiscate) without noise resulting from  $\hat{J}$  (going out) because both are written as  $\hat{J}$  when not diacritized. Thus, expert users would prefer that word sketches be computed on diacritized forms. The provisional solution is two versions of the corpus: one for users who know they need to use diacritized forms to obtain word sketches, the other for those who do not. We are currently building an interface option that allows users to use the undiacritized form while keeping the diacritized form as an option for advanced users.

We must note here that the quality of the output of the system depends heavily on the input, i.e., the quality of tagging and lemmatization. Errors in lemmatization and tagging will not go unnoticed and can lead to unexpected results for the lexicographer. There is generally a logical explanation, but it may require a closer view into the tagging and lemmatization to fully understand the output. One general difficulty is with proper nouns whose form is ambiguous with another word. For example, the name  $q_{ab}$  (Huyay) is a common first name in religious texts. However, MADA usually tags it as an adjective meaning "modest", a mistake that stems from the fact that MADA is mostly built to process modern standard Arabic (MSA) texts, where this name is not a common one. It is also assigned the wrong lemma:  $q_{ab}$  (Huyiy~) instead of  $q_{ab}$  (Huyay~). Thus, when the lexicographer wants to search for words that may be read as proper nouns or adjectives, they must be aware of the ambiguity and either use the wrong lemma or search only with the simple string.

On the results page, the concordances are shown, by default, in a keyword-in-context (KWIC) view, as in Figure 2. With VIEW options, it is possible to change the concordance view to a number of alternative views. One is to view additional attributes such as POS tags or lemma alongside each word. This can be useful for finding out why an unexpected corpus line has matched a query, e.g., because of an incorrect POS-tag or lemma. By selecting fields in the references column, the user can decide what source of information should appear in blue at the left-hand end of the concordance line.

### 5 Summary and future plans

We have presented arTenTen, a very large web-crawled corpus of contemporary Arabic. We have also presented in some detail the subset of that corpus that has been processed by the MADA tool: how it has been set up and encoded and how we have produced word sketches for Arabic, with a full account of the sketch grammar that was used. We have discussed how this MADA-processed corpus can be used for dictionary-editing and related linguistic research, including how it can be used to find collocations, idioms, new words, new senses, and via the thesaurus, synonyms and related words. We have introduced the sketch diff, which shows how near-synonyms can be compared and contrasted.

We would of course like to apply MADA to the whole of arTenTen. To date, this has not been possible because of the speed of the program. This has recently been addressed with MADAMIRA (Pasha *et al.* 2014), a new and improved version of MADA combined with AMIRA (Diab 2009) that is orders of magnitude faster than MADA and has an output of comparable quality.

The method of compilation of arTenTen aims at a diverse corpus, including texts from many domains and genres. The nature of the Arabic language family also means that web texts are likely to appear in many language varieties: modern standard Arabic (MSA), classical Arabic, Quranic Arabic, and various dialects. Identifying the language variety of each text (or sub-text unit) is thus both a challenge and an opportunity: it is a non-trivial task, although standard language identification methods work quite well on identifying Arabic dialects (Zaidan and Callison-Burch 2013). The opportunity that lies in identifying the language varieties will facilitate lexicographic work on specific varieties and the comparative study of the dialects.

In preliminary experiments, we built a classifier to distinguish between MSA, classical Arabic, and Egyptian, Jordanian, and Saudi dialects. We trained a five-gram character level language model for each of these varieties based on published corpora and tested its performance on a small, manually selected subset of arTenTen texts in MSA, classical Arabic, and Egyptian Arabic, achieving 93% accuracy in this three-wise classification task. Then, we trained a combined dialectal model based on the Egyptian, Jordanian, and Saudi texts and processed a large number of arTenTen texts (40 k). We observed that the majority of the texts (~80%) are identified as MSA, and the rest are identified as classical or dialectal Arabic. This shows that a non-negligible portion of the texts is non-MSA. In future work, we intend to improve our language variety identification and increase its coverage to other dialects, using corpus-based approaches and resources, such as Buckwalter and Parkinson's Frequency Dictionary (2011) and the keywords method presented in Kilgarriff (2012). We will also consider the identification of sub-text units (Elfardy and Diab 2013), which is important for mixed texts.

arTenTen was gathered in 2012; so, it is already two years old. For each of the TenTen corpora, a program of re-crawling is planned, whereby material will regularly be added, both to keep the corpus current and so that empirical methods can be applied to the discovery of new words and meanings. We intend to gather newspaper feeds and blog feeds so that we have additional material with accurate time stamps.

We believe arTenTen, in combination with MADA/MADAMIRA and the Sketch Engine, possesses considerable promise for improved Arabic linguistic description and lexicography.

### Acknowledgments

This work was partly supported by the Ministry of Education of the Czech Republic within the LINDAT-Clarin project LM2010013 and by the Ministry of the Interior of the Czech Republic within the project VF20102014003. Nizar Habash performed most of his work on this article while he was at the Center for Computational Learning Systems at Columbia University.

# References

Alansary, S., Nagi, M. and Adly, N. (2007) Building an International Corpus of Arabic (ICA): progress of compilation stage. In: *7th International Conference on Language Engineering*, Cairo, Egypt.

Alkuhlani, S. and Habash, N. (2011) A Corpus for Modeling Morpho-Syntactic Agreement in Arabic: Gender, Number and Rationality. In: *Proceedings of the Association for Computational Linguistics (ACL'11)*, Portland, Oregon.

Alrabiah, M., Al-Salman, A. and Atwell, E. (2013). The design and construction of the 50 million words KSUCCA King

Saud University Corpus of Classical Arabic. In: Second Workshop on Arabic Corpus Linguistics (WACL-2), Lancaster, UK 2013.

- Al-Sulaiti, L. and Atwell, E. (2006) The design of a corpus of Contemporary Arabic. *International Journal of Corpus Linguistics*, **11** (2).
- Arts, T. et al. (Forthcoming 2014) Oxford Arabic Dictionary. Oxford University Press.
- Buckwalter, T. Buckwalter Arabic Morphological Analyzer v2.0. LDC Catalog No.: LDC2004L02. Linguistic Data Consortium.
- Buckwalter, T. and Parkinson, D. (2011) A Frequency Dictionary of Arabic. Routledge. Frequency Dictionary Series.
- Diab, M. (2009) Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In: *2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.
- Eckart, T., Quasthoff, U., Alshargi, F. and Goldhahn, D. (2014) Large Arabic Web Corpora of High Quality: The Dimensions Time and Origin. In: Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools, LREC 2014. Reykjavik, Iceland.
- Elfardy, H. and Diab, M. (2013) Sentence Level Dialect Identification in Arabic. In: Proceedings of the Association for Computational Linguistics (ACL'13), Sofia, Bulgaria.
- Graff, D. (2003) Arabic Gigaword. LDC Catalog No.: LDC2003T12. Linguistic Data Consortium.
- Graff, D., Maamouri, M., Bouziri, B., Krouna, S., Kulick, S. and Buckwalter, T. (2009) Standard Arabic Morphological Analyzer (SAMA) Version 3.1. LDC Catalog No.: LDC2009E73. Linguistic Data Consortium.
- Habash, N. (2010) Introduction *to Arabic Natural Language Processing*. Morgan & Claypool Publishers. Synthesis Lectures on Human Language Technologies.
- Habash, N. and Rambow, O. (2005) Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In: *Proceedings of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan.
- Habash, N., Rambow, O. and Roth, R. (2009) MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In: Proceedings of the International Conference on Arabic Language Resources and Tools, Cairo, Egypt.
- Habash, N., Roth, R., Rambow, O., Eskander, R. and Tomeh, N. (2013) Morphological Analysis and Disambiguation for Dialectal Arabic. In: *Proceedings of Conference of the North American Association for Computational Linguistics (NAACL)*, Atlanta, Georgia.
- Habash, N., Soudi, A. and Buckwalter, T. (2007) On Arabic Transliteration. In: A. van den Bosch and A. Soudi (eds.). *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.
- Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. and Suchomel, V. (2013) The TenTen Corpus Family. In: *International Conference on Corpus Linguistics*, Lancaster, UK.
- Kilgarriff, A. (2012) Getting to Know your Corpus. In: *Proceedings of Text, Speech, Dialogue Conference*, Brno, Czech Republic 2012. Springer.
- Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. (2004) The Sketch Engine. In: *Proceedings of EURALEX*, Lorient, France 2004. pp 105-116.
- Maamouri, M., Bies, A., Buckwalter, T. and Mekki, W. (2004) The Penn Arabic Treebank : Building a Large-Scale Annotated Arabic Corpus. In: *NEMLAR Conference on Arabic Language Resources and Tools*, Cairo, Egypt 2004. pp 102–109.
- Pasha, A., Al-Badrashiny, M., El Kholy, A., Eskander, M., Diab, M., Habash, N. Pooleery, M., Rambow, O. and Roth, R (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland.
- Pomikalek, J. (2011) Removing Boilerplate and Duplicate Content from Web Corpora. PhD, Masaryk University.
- Pomikalek, J. and Suchomel, V. (2012) Efficient web crawling for large text corpora. In: *Proceedings of the* 7<sup>th</sup> Web as *Corpus Workshop (WAC7)*, Lyon, France.
- Roth, R., Rambow, O., Habash, N., Diab, M. and. Rudin, C. (2008) Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In: *ACL 2008: The Con ference of the Association for Computational Linguistics Companion Volume, Short Papers*, Columbus, Ohio.
- Rychlý, P. (2008). A lexicographer-friendly association score. In: *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008,* Karlova Studanka, Czech Republic.
- Sharoff, S. (2006) Creating general-purpose corpora using automated search engine queries. In: M. Baroni & S. Bernardini (eds.). *WaCky! Working papers on the Web as Corpus*. Bologna, Italy.
- Wehr, H. (1979) Dictionary of Modern Written Arabic 4th ed., Spoken Language Services.
- Zaidan, O. and Callison-Burch, C. (2013) Arabic Dialect Identification. Computational Linguistics.
- Zaghouani, W. (2014) Critical Survey of the Freely Available Arabic Corpora. In: *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools, LREC 2014.* Reykjavik, Iceland.

### **Appendix 1: Arabic Sketch Grammar**

```
# arTenTen Sketch Grammar, version 0.1 (7/20/2013)
*STRUCTLIMIT s
*DEFAULTATTR tag
*FIXORDER subject/subject-of adjective/adjective-of construct-state and/or
*DUAL
=subject/subject-of
        1:"verb" 2:[tag="noun" & case="n" & pref1tag!="prep" & pref2tag!="conj"]
*DUAL
=adjective/adjective-of
        1:"noun" 2:[tag="adj" & pref1tag!="prep" & pref2tag!="conj"] & 1.state = 2.state & 1.case = 2.case
        1:"noun" [tag="adj" & pref1tag!="prep" & pref2tag!="conj"] 2:[tag="adj" & pref1tag!="prep"] & 1.state =
2.state & 1.case = 2.case
# noun adjective pair; enforce agreement in state and case
=construct-state
        1:[tag="noun" & state="c"] 2:[tag="noun" & case="g" & pref1tag!="prep" & pref2tag!="conj"]
# simple annexation
#1:[tag="noun" & state="c"] [tag="noun" & case="g" & state="c" & pref1tag!="prep" & pref2tag!="conj"]+
[tag="noun" & case="g" & pref1tag!="prep" & pref2tag!="conj"]
# more complex annexation
=and/or
*SYMMETRIC
        1:"noun" [trans=">w"|trans=">m"|trans="w"] 2:"noun" & 1.state = 2.state & 1.case = 2.case
        1:"noun" 2:[tag="noun" & pref2="wa"] & 1.state = 2.state & 1.case = 2.case
# noun
        1:"adj" [trans=">w"|trans=">m"|trans="w"] 2:"adj" & 1.state = 2.state & 1.case = 2.case
        1:"adj" 2:[tag="adj" & pref2="wa"] & 1.state = 2.state & 1.case = 2.case
# adjective
        1:"verb" [trans=">w"|trans=">m"|trans="w"] 2:"verb" & 1.aspect = 2.aspect
        1:"verb" 2:[tag="verb" & pref2="wa"] & 1.aspect = 2.aspect
# verb
```

### Appendix 2: Sample arTenTen XML 'vertical' format

With selected attributes of a morphological annotation by MADA. There are two paragraphs () each with one sentence (<s>) within one document (<doc>). The source of the document and other metadata is stored in attributes of structures (e.g. url="http://www.alsabar-mag.com/ar/article 419").

	oc id="301	<b>n diac</b> " length="66	lemma voc latin 15" url="ht	a voc	l <b>emma</b> latin /www.al	<b>lemma</b> sabar-		<b>tag</b> ar/arti	bw ccle419">	pers	son asj	pect	voice	mood	l gender	number	state	case	gloss	lex/ punc
	id="8135"	>																		
مات	klmAt کا	kalimAti	kalimap_1	كَلِمَة	klmp	كلمة	kalim	noun	+kalim/NOUN+At/NSUFF_F EM_PL+i/CASE_DEF_ACC li/PREP+Al/DET+baHov/NO						f	р	С	a	words;remarks	lex
حث ۱۶/> ۱/>	;>	lilbaHovi	baHov_1	بَحْث	bHv	بحث	baHov	noun	UN+i/CASE_DEF_GEN						m	S	d	g	discussion	lex
	•																			
<s< td=""><td>id="8136"</td><td>&gt;</td><td></td><td></td><td></td><td></td><td></td><td></td><td>Al/DET+nASir/ADJ+ap/NSU</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></s<>	id="8136"	>							Al/DET+nASir/ADJ+ap/NSU											
اصرة	AlnASrp ال	Aln~ASirapi	nASir_2	ناصِر	nASr	ناصر	nASir	adj	F_FEM_SG+i/CASE_DEF_G N	r£					f	S	d	g	partisan;supporter	lex
:	:	:	:_0	:	:	:		punc	:/PUNC +inoTibAE/NOUN+At/NSUF										:	punc
عات	AnTbAE انطبا.	At AinoTibAEAtl	N {inoTibAE_1	ٱِنْطِباع	{nTbAE	أنطباع	{inoTibAE	noun	_FEM_PL+N/CASE_INDEF_ OM	_N					f	р	i	n	impression	lex
من	mn	min	min_1	مِن	Mn	من	min	prep	+min/PREP+ Al/DET+barolamAn/NOUN+i	;/									from	lex
لمان	AlbrlmAı البر	n AlbarolamAni	barolamAn_1	بَرْلَمان	brlmAn	برلمان	barolamAn	noun	CASE_DEF_GEN	1/					m	S	d	g	parliament which;who;whom	lex
الني	Al*y	Al~a*iy	Al~a*iy_1	الَّنِي	Al*y	الذي	Al~a*iy	pron_rel	+Al~a*iy/REL_PRON+						m	S	i	u	[masc.sg.] be_held;be_con-	lex
عقد	Eqd	Euqida	Eaqad-i_1	عَقَد	Eqd	عقد	Euqid	verb	+Euqid/PV_PASS+a/PVSUFF SUBJ:3MS	7_ 3	р	I	р	i	m	S			vened;be_con- cluded	lex
في	fy	fiy	fiy_1	فِي	Fy	في	fiy	prep	+fiy/PREP+ +Hadiyq/NOUN+ap/NSUFF_										in	lex
ديقة	> Hdyqp	HadiyqapK	Hadiyqap_1	حَدِيقَة	Hdyqp	حديقة	Hadiyq	noun	EM_SG+K/CASE_INDEF_G						f	S	i	g	garden	lex
عامة ؟/>	1	EAm~apK	EAm~_1	عامّ	EAm	عام	EAm~	adj	+EAm~/ADJ+ap/NSUFF_FEN _SG+K/CASE_INDEF_GEN						f	S	i	g	general;common;j ublic	lex
1</td <td></td> <td>6 H</td> <td></td>		6 H																		

More paragraphs follow... </doc>