arTenTen: Arabic Corpus and Word Sketches

Tressy Arts
Lexicographer
tressy.arts@gmail.com

Yonatan Belinkov
MIT, USA
belinkov@mit.edu

Nizar Habash
New York University Abu Dhabi, UAE
nizar.habash@nyu.edu

Adam Kilgarriff
Lexical Computing Ltd, UK
adam@lexmasterclass.com

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 5.8-billion 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.[1] 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.[2] 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.[3]

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

1 Arabic Gigaword is created and distributed by the Linguistic Data Consortium (Graff 2003). It is regularly updated and is now in its fifth edition.
2 http://www.sketchengine.co.uk
3 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,[4] Estonian, Irish,[5] 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 طفل (child) searches for the lemma as well as the string; so, the strings الطفل (the+child), طفلهما (child+their), كالأطفال (like+the+children), etc., are all retrieved.

Figure 1: Simple concordance query

4 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.

5 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.

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[6] forms function on the left hand menu (Figure 3) shows which of the returned forms are most frequent.

6 The Nodes are the concordance result, i.e. all tokens from the corpus matching the concordance query.

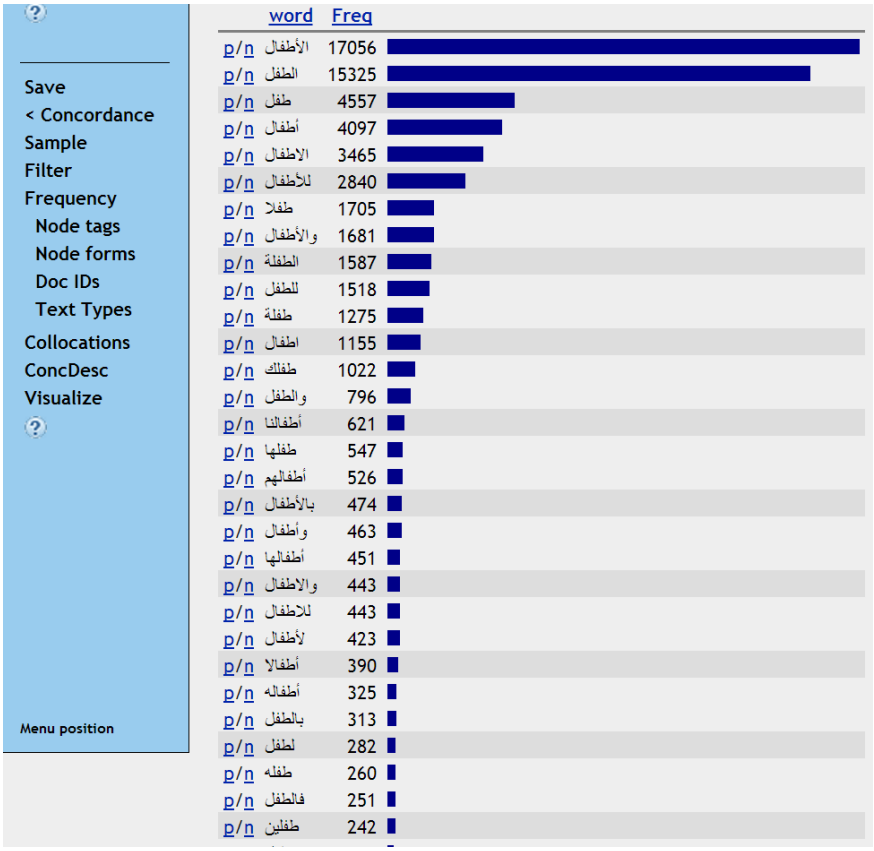

Figure 3: Frequency of node forms of طفل

The **p**/**n** links are for positive and negative examples. Clicking on **p** gives a concordance for the word form, while clicking on **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).

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

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

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 طفل is frequent on *.edu* sites.

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

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

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[7] lemmas for the corpus.

Word list

Corpus: **arTenTen12 [sample 115M]**

Page 1 Go Next >

word	Freq
في	3242280
من	2914934
على	1593477
أن	1184760
إلى	754664
عن	738288
لا	659851
و	637527
الله	629086
ما	610949
التي	585503
هذا	518842
أو	453099
الذي	416753
ان	413353
مع	402313
هذه	402083
كان	361499

Word list

Corpus: **arTenTen12 [sample 115M]**

Page 1 Go Next >

lemma	Freq
في	3962066
من	3500214
على	2285678
أن	2184612
الذي	1310294
هذا	1245137
إلى	1231294
كان	1102480
ما	1009041
لا	984894
عن	927882
إن	899877
قال	746755
الله	722219
ذلك	640003
و	638513
-ل	588106
أو	545702

Word list

Corpus: **arTenTen12 [sample 115M]**

Page 1 Go Next >

lemma_voc	Freq
فِي	3962066
مِن	3413373
عَلَى	2283548
أَنْ	1332439
الَّذِي	1310294
هَذَا	1245137
إِلَى	1231294
كان	1102480
ما	1009041
لا	984894
عَن	927455
أَنَّ	850948
قال	746755
الله	722219
ذَلِكَ	640003
وَ	638513
-لِ	588106
إِنْ	557135

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 لون (color) but also the idiomatic الأخضر واليابس (literally "the green and the dry"). Clicking on the number after the collocate gives a concordance of the combination (Figure 10).

7 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.

أخضر

arTenTen12 [sample 115M] freq =

and/or	472	0.7
يابس	174	12.6
أصفر	52	9.52
برتقالي	9	8.45
أزرق	13	8.31
بنفسجي	5	8.05
وردي	6	7.88
أحمر	77	7.87
رمادي	4	7.19
أبيض	42	6.99
جاف	5	6.78
أسود	12	6.36
ممتاز	6	5.72
نظري	4	5.13
مفتوح	4	4.48

adjective-of	4865	4.6
ضوء	456	10.48
شاي	178	10.14
لون	300	9.42
خط	355	8.94
مسطح	61	8.54
منطقة	637	8.41
جبل	119	8.35
حزام	57	8.23
عشب	51	8.1
مسيرة	83	7.76
مساح	67	7.65
نبات	38	7.25
علف	26	7.24
شجرة	48	7.23
فلفل	22	7.12
بصل	21	6.95
جزيرة	57	6.94
رقعة	22	6.87
وادي	21	6.86
قبة	23	6.83
راية	25	6.8
ورق	54	6.7

Page 1 of 9 Go Next | Last

http://www...	الأراضي اللبنانية ذ 5 كانون الأول : حرائق هائلة تقضي على **الأخضر واليابس** في كل لبنان والمحكمة الدولية تنفي تقريرا
http://isl...	العولمة , الإمبراطور الأخير للحداثة المسيحية , ستأتي على **الأخضر واليابس** في تربتنا وفي تربتهم , عدا في حالة واحدة
http://www...	والمجتمع والجماعات والأفراد في تصاعد وتيرته حتى أكل **الأخضر واليابس** ? ولا شك أن الظاهرة الاحتجاجية تدفعنا إلى
http://www...	مجتمعية ممكنة اتقاء للهزات والعواصف التي تأتي على **الأخضر واليابس** , كما حدث فعلا في بعض الأقطار العربية في
http://www...	ان هذه القلعة هى المصدر الوحيد للرزق لكنهم قضوا على **الاخضر واليابس** ..... فقد بيعت هذه الشركة بمبلغ 1.3 مليار
http://hem...	وجاهليته وحقارته إذا كان الهدف فتح حرب الطوائف لتحرق **الأخضر واليابس** . لكننا لا نعرف بالتحديد من يقف وراء العملية
http://www...	" الوهم الطويل مدو ومرعب حتى لأعدائه , هو الذي أدمن **الأخضر واليابسة** " في مرحلة صعوده , يأبى إلا أن يطالهما وهو
http://tun...	القرى والمدن , وجميعنا يعلم أن ظلم نظام بن علي أتى على **الأخضر واليابس** من بنزرت لبرج الخضراء , ولكن أن تصبح شاشاتنا
http://www...	سكان عدد كبير من القرى إذ قضت عليها تماما وأتت على **الأخضر** هل المصائب التي . ? **واليابس** معا في ميانمار
http://www...	والاتفاق تترافق مع دعوات الى المواجهة والحرب وحرق **الأخضر واليابس** وترويع الناس ? وأي نظام أمني رسمي يمكن أن
http://ksa...	حاق بفرص عمل ثمينة فارتموا في احضان مصارف قضت على **الاخضر** وبكثير من الامتعاض تقبل . **واليابس** في جيوبهم
http://www...	ومن يتبعه من المتعودين على الولائم والبزنس بدرس وطحن **الأخضر واليابس** الذي لا يمت بصلة لمشاريعهم بالاحتيال وتقاسم
http://ndb...	اليوم فلا حيلة لي سوى الصمت فبواخرهم اصبحت تاتي على **الاخضر واليابس** وانا رغم ماقدمت لازلت انا انا لا مرافق ولا
http://www...	سبعة أعوام عجاف أكل المستعمر وعملائه المظلوم **الأخضر واليابس** في عراقنا الجريح ودمر جميع مؤسسات الدولة
http://aln...	حقيقي لها , بل هي مسمى فقط , وإن هي الا فوضى أتت على **الأخضر** . **واليابس** أضاعت وحدة البلاد واستقرارها وحريتها
http://www...	القوى المختلفة , بل قد تشهد مصر حربا أهلية تأتي على **الأخضر واليابس** , وفي ظل هذه المعطيات فإن من يدعو إلى إسقاط
http://dai...	وان نوأد هذه الفتن فى مهدها قبل ان تتفشى وتقضى على **الاخضر** .. **واليابس** وتكون العواقب وخيمة و يكون قد فات الوقت
http://www...	استفاقت مدينة ليون الفرنسية على حريق أتى على **الأخضر واليابس** في مستودع كبير للحافلات مملوك لشركة كيوليس
http://zef...	البدرى ولعبه بطريقه لعب لاتناسب الفريق كادت ان تاتى على **الاخضر واليابس** فى الفريق وتسببت بالفعل فى خروج الفريق من
http://www...	واحتدمت الأمور وانفجرت الحرب الأهلية التي أتت على **الأخضر** يغادر أعضاء الفريق البرتقالي مدينة . **واليابس**

Page 1 of 9 Go Next | Last

Figure 9: Word Sketch results for أخضر (left)

Figure 10: Concordance lines for أخضر in combination with its collocate يابس (right)

In this concordance, we see that this combination usually occurs with أتى على (10 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 الأخضر واليابس and the idiom أتى على الأخضر واليابس (to destroy everything).

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

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

Figure 11: Concordance lines for أخضر in combination with ضوء

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*.

Some of the same things are أحمر/red in Arabic and English; thus, we find the matched pairs لحم/meat, سجاد/carpet, and فلفل/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.

adjective-of	9782	4.8	modifies	778081	0.4
صليب	1132	11.52	flag	27161	9.32
هلال	1101	11.38	carpet	21070	9.04
بحر	1498	11.0	wine	34956	8.8
خط	845	9.92	tape	15035	8.44
لون	507	9.75	meat	17594	8.34
قلعة	214	9.19	pepper	10687	8.26
بطاقة	249	9.08	herring	6067	7.93
لحم	208	8.87	light	25976	7.58
هندي	157	8.6	onion	5258	7.33
سافية	93	8.24	rose	4714	7.3
دم	212	7.95	cell	16369	7.22
كرية	75	7.93	lipstick	3047	6.86
سجاد	77	7.92	ink	3605	6.73
شمع	67	7.76	bump	2858	6.59
شيطان	87	7.7	grape	2877	6.51
زاوية	69	7.54	stripe	2462	6.43
ياقوت	47	7.23	ribbon	2586	6.41
فلفل	46	7.22	sole	2300	6.39
ورد	58	7.07	lip	3592	6.35
طوب	34	6.78	shirt	4524	6.33
زاءوق	32	6.72	berry	2563	6.31
كبريت	32	6.71	hair	9863	6.29
خمير	32	6.71	dress	6136	6.24
درب	34	6.64	snapper	1856	6.24
بقعة	33	6.62	arrow	2207	6.19

Figure 12: Adjective results of a bilingual word sketch for Arabic أحمر and English *red*

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 استيراد (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.

تصدير arTenTe

Lemma	Score	Freq
أستيراد	0.459	3611
تسويق	0.343	5379
تخزين	0.322	2285
تصنيع	0.307	3483
أستهلاك	0.3	5377
توريد	0.287	1190
أستخراج	0.266	2456
شراء	0.265	15838
تهريب	0.254	3817
تجارة	0.251	17570
إنتاج	0.248	28181
تجميع	0.244	2534
توزيع	0.238	16139
ترويج	0.237	4061
إدخال	0.235	6883
بيع	0.226	28032
تداول	0.221	9957
أستبدال	0.197	3251
نوعية	0.196	4059
تدفق	0.194	3801
تفريغ	0.189	1618

Figure 13: Thesaurus search showing entries similar to تصدير (export) (left)

تصدير / أستيراد arTenTen12 [sample 115M] freqs = 5037 /3611

تصدير 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 أستيراد

subject-of	289	252	1.3	1.7
حظر	3	8	4.8	6.3
جاز	3	5	3.8	4.5
يتم	48	66	6.4	6.9
تم	74	99	4.9	5.3
تام	10	9	4.2	4.1
جرى	7	5	4.5	4.0
كان	10	0	0.6	--
مكن	15	8	4.0	3.1
منع	11	5	5.5	4.3
أستطاع	3	0	3.0	--
وقف	4	0	3.3	--
أراد	5	0	3.9	--
بدأ	9	0	4.0	--
توقف	3	0	4.1	--
أمكن	4	0	4.2	--
حاول	7	0	4.9	--
أعاد	4	0	5.7	--
أتاح	3	0	6.6	--
أوقف	6	0	7.6	--
أحتكر	3	0	7.6	--

adjective	141	129	0.2	0.3
ترفي	0	3	--	9.0
عشوائي	0	13	--	7.4
أستهلاكي	0	5	--	6.9

construct-state	1892	1129	2.0	1.9
مشتق	0	12	--	7.5
خضار	0	9	--	7.1
بنزين	0	8	--	6.8
خضرة	0	8	--	6.7
لباس	0	7	--	6.7
لحم	9	52	5.5	8.2
دواء	8	40	4.9	7.3
زيت	6	18	5.3	7.1
قمح	14	37	7.0	8.7
سلع	18	41	6.5	7.9
كمية	22	33	6.3	7.0
سلعة	8	8	6.2	6.5
أسمنت	20	18	7.6	7.8
بضاعة	29	25	7.5	7.5
أرز	15	7	7.2	6.4
منتوج	10	3	6.8	5.5
سلاح	83	26	6.8	5.1
منتج	84	24	7.7	6.0
نفط	219	39	8.7	6.2
غاز	443	52	10.3	7.3
بترول	36	3	7.7	4.4
إرهاب	42	0	6.6	--
ثورة	134	0	6.8	--
شحنة	10	0	6.9	--

and/or	602	412	1.6	1.6
تصدير	0	261	--	11.5
جمرك	0	3	--	5.5
شراء	0	5	--	4.9
تداول	0	3	--	4.0
توزيع	5	15	4.3	5.9
أستهلاك	10	6	6.3	5.6
بيع	11	6	4.9	4.1
نقل	19	10	4.6	3.7
صناعة	9	3	4.3	2.7
تسويق	18	4	7.1	5.0
إنتاج	63	14	6.7	4.6
سعر	3	0	2.3	--
تجارة	7	0	3.7	--
أستثمار	18	0	5.3	--
تحميل	3	0	5.4	--
تحضير	4	0	5.6	--
تخزين	4	0	6.2	--
شحن	9	0	6.8	--
توريد	4	0	6.8	--
تكرير	5	0	7.4	--
تصنيع	17	0	7.7	--
أستخراج	9	0	7.7	--
أستيراد	261	0	11.7	--

Figure 14: Sketch Diff comparing collocates of تصدير 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.

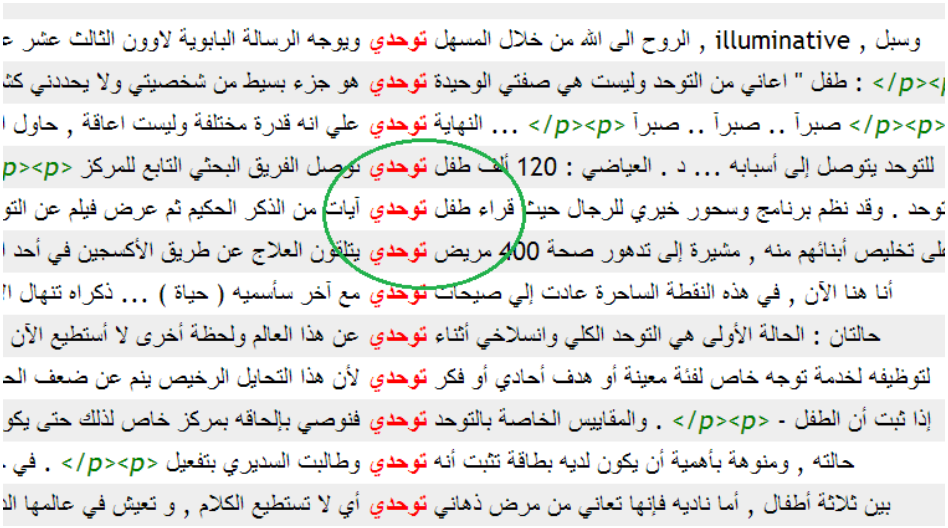

Figure 15: Concordance for توحي

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.

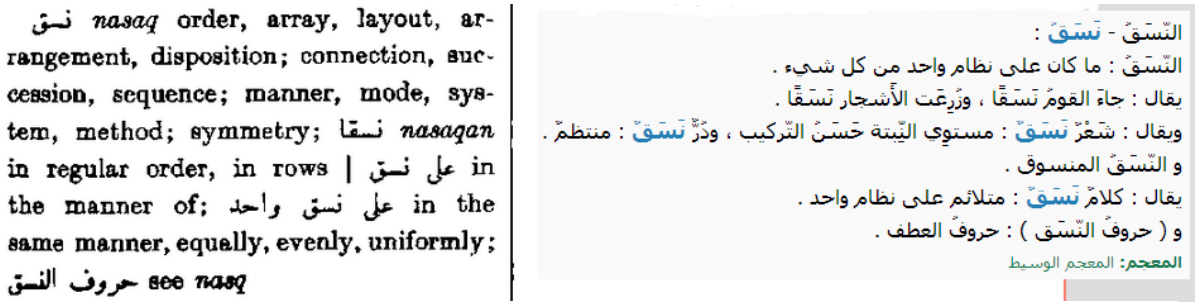

Figure 16: Dictionary entries for نسق from Wehr's *Dictionary of Modern Written Arabic* 4th ed. 1979, and *al-mu'jam al-wasit* (Academy of the Arabic Language in Cairo)[8]

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

Figure 17: Concordance for نسق with تصاعي

8 Entry as found at almaany.com, February 2014.

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

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., نسق,تصاعي) 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[9] 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.[10] 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.

9 The association score currently in use is a variant of the Dice coefficient; see Rychlý (2008) for full details.
10 See also section 4.2.

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*.[11]

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 المملكة 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).

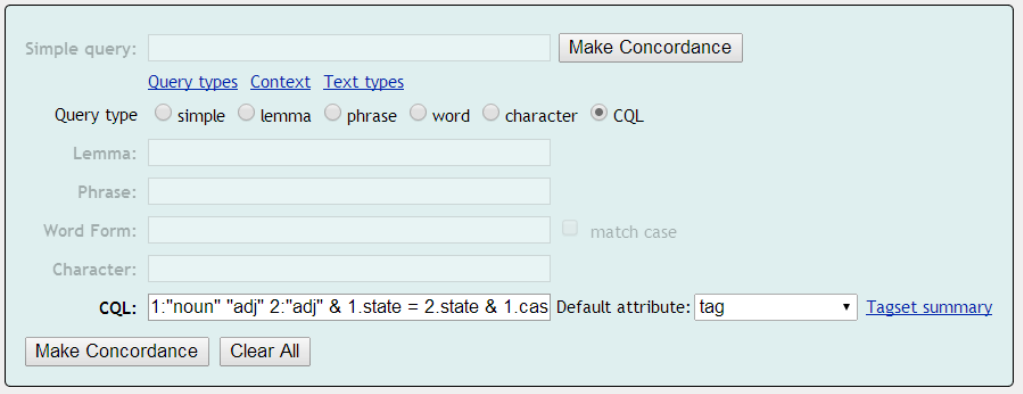

Figure 18: Using CQL in the concordance search form (with *tag* as default attribute)

11 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).

Query noun, adj 25 > Random sample 25 (0.2 per million)

Page 1 of 2 Go Next | Last

307351	, من قبل عملاء النظام في المنطقة نفسها	الأجواء القمعية المشددة	في إيران . وفي الليلة التي سبقتها ورغم
417451	فاذا لم توجد افعال حقيقية من قبل واشنطن .	المشروع الذري الايراني	وواشنطن ستنفذ كل الخطوات الضرورية لوقف
544001	في البلاد .. استيقظ من سباتك !! سنة و 9	والاختلالات السياسية والامنية	تسمع عن الفساد والفوضى والحرب في الصعده
666351	وذكرت مصادر صحفية ان وزير التربية .)	الانسان العراقي الجديد	وتطويرها تحت شعار (نحو منهج تربوي متطور لبناء
862751	هذا دليل على اننا امة 23/12/2008 ",	المملكة العربية السعودية	" , عبدالرحمن محمد !! وهذا واحد منهم
1121401	لسوريا لتنفيذ هجمات على اجهزة الامن في	نائب لبناني مناهض	رجال قالوا انهم تلقوا أموالا وأسلحة من
2150051	للمنطقة الحرة , والتي أبرم بعضها في عام	المصانع المحلية التابعة	قسرا بما تقول فما الحال إذن في عقود منح
2821501	مو لهدف التجريح مو لهدف ان ما نزعل أحد	مراجعة دقيقة وصريحة	هالفكرة عفا عليها الزمن علينا مراجعتها
3232751	وبين ترات هذا الاجتماع , ان هذا الاجتماع	للاجتماع العربي الاسلامي	لرصد العلاقة بين ارهاصات التفكير بحداثة
3559851	ودستور الأغذية العالمي . ولخطورة وجود هذا	المواصفات السعودية والأمريكية	تحتوي على نسبة تفوق الحد المسموح به من قبل
4803851	وعدم الإنصياع وراء نصيحة قريب بدت عليه ,	لأزمة صحية حقيقية	فكم من عارض صحي يبدو تافها يكون مؤشرا ,
4816301	, لا يقودها هذه المرة ريتشارد قلب الأسد	حرب صليبية جديدة	الموجهة إلى العراق في شهر رمضان على أنها
5737151	جديدة يعتمد فيها على مقدمة قانون التوزن	برؤية علمية وعصرية	ـ قانون الاعمار وهو قانون 10 . الانتخابات
6352651	وفرنسا مع أصدقائها اللبنانيين ستدفع في ,	الأيام القليلة المقبلة	نبيه بري اعتقاده أن " الأمور ستتبلور خلال
6510601	للتصنيفات التي يفترضها التأريخ أو النقد	البحث المنهجي الملائم	فائزين في هذا العمل , ولا أي شكل من أشكال
8194451	2011-11-09 دورو على الموظفين الحيتان	انسان متواضع ونظيف	ليثر كل الهجوم على معاليه مع انه للامانه
8244451	في تفاقم هذه القضية هو أن المعتدي غالبا	السبب المساند والمؤثر	وأوضح أن . بما فيها الرغبات الشاذة

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 (the rain fell) produces two grammatical relations. When نزل (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 نزل.

- **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 بحث takes the *adjective* علمي, which itself is *adjective-of* for بحث.

- **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 كبير (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 يبكي, "laughs or cries").

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[12] 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[13] (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

12 http://www.sketchengine.co.uk/documentation
13 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[14]

Table 1: Data sizes at the various stages of corpus preparation

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.com	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 وبفكرة *wbfkrp* assuming a specific analysis corresponding to the English 'and with an idea'.

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

prc1:bi_prep	First proclitic position away from base word حرف \ سابقة جر
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	Voice (not applicable here) (البناء (معلوم/مجهول
mod:na	Mood (not applicable here) الصيغة
gen:f	Gender (feminine here) الجنس
num:s	Number (singular here) العدد
stt:i	State (indefinite here) التعريف
cas:g	Case (genitive here) الحالة لإعرابية
enc0:0	Only enclitic after the base word ضمير \لاحقة متصل
spvar:lex	Spelling Variant (none, exact lexicon match here) إملاء غير قياسي
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 باتمان؟

Will **Ben** Affleck be a good Batman?

(2) كيري يحاول مجددا انقاذ المفاوضات **بين** فلسطين واسرائيل

Kerry tries again to save the negotiations **between** 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 صَادَر (to confiscate) without noise resulting from صَادِر (going out) because both are written as صادر 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 حيي (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: حَيِيّ (Hayiy~) instead of حُيَيّ (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 7th 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 (`<p>`) 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"`).

word	word latin	diac	lemma voc latin	lemma voc	lemma latin	lemma	stem	tag	bw	person	aspect	voice	mood	gender	number	state	case	gloss	lex/ punc
<doc id="301" length="6615" url="http://www.alsabar-mag.com/ar/article__419">																			
<s id="8135">																			
كلمات	klmAt	kalimAti	kalimap_1	كَلِمَة	klmp	كلمة	kalim	noun	+kalim/NOUN+At/NSUFF_FEM_PL+i/CASE_DEF_ACC					f	p	c	a	words;remarks	lex
للبحث	llbHv	lilbaHovi	baHov_1	بَحْث	bHv	بحث	baHov	noun	li/PREP+Al/DET+baHov/NOUN+i/CASE_DEF_GEN					m	s	d	g	discussion	lex
</s>																			
<s id="8136">																			
الناصرة	AlnASrp	Aln~ASirapi	nASir_2	نَاصِر	nASr	ناصر	nASir	adj	Al/DET+nASir/ADJ+ap/NSUFF_FEM_SG+i/CASE_DEF_GEN					f	s	d	g	partisan;supporter	lex
:	:	:	:_0	:	:	:		punc	:/PUNC									:	punc
انطباعات	AnTbAEAt	AinoTibAEAtN	{inoTibAE_1	إِنْطِبَاع	{nTbAE	آنطباع	{inoTibAE	noun	+inoTibAE/NOUN+At/NSUFF_FEM_PL+N/CASE_INDEF_NOM					f	p	i	n	impression	lex
من	mn	min	min_1	مِن	Mn	من	min	prep	+min/PREP+									from	lex
البرلمان	AlbrlmAn	AlbarolamAni	barolamAn_1	بَرْلَمان	brlmAn	برلمان	barolamAn	noun	Al/DET+barolamAn/NOUN+i/CASE_DEF_GEN					m	s	d	g	parliament	lex
التي	Al*y	Al~a*iy	Al~a*iy_1	الَّتِي	Al*y	التي	Al~a*iy	pron_rel	+Al~a*iy/REL_PRON+					m	s	i	u	which;who;whom_ [masc.sg.]	lex
عقد	Eqd	Euqida	Eaqad-i_1	عَقَد	Eqd	عقد	Euqid	verb	+Euqid/PV_PASS+a/PVSUFF_SUBJ:3MS	3	p	p	i	m	s			be_held;be_convened;be_concluded	lex
في	fy	fiy	fiy_1	فِي	Fy	في	fiy	prep	+fiy/PREP+									in	lex
حديقة	Hdyqp	HadiyqapK	Hadiyqap_1	حَدِيقَة	Hdyqp	حديقة	Hadiyq	noun	+Hadiyq/NOUN+ap/NSUFF_FEM_SG+K/CASE_INDEF_GEN					f	s	i	g	garden	lex
عامة	EAmp	EAm~apK	EAm~_1	عامّ	EAm	عام	EAm~	adj	+EAm~/ADJ+ap/NSUFF_FEM_SG+K/CASE_INDEF_GEN					f	s	i	g	general;common;public	lex
</s>																			
More paragraphs follow...																			
</doc>																			