Corpus-Based Prediction of Coordination Ambiguity in Arabic

Wafaa Daffa, Raad Alshahry, Imtiaz Hussain Khan
Department of Computer Science, King Abdulaziz University
Jeddah, Saudi Arabia
wafaa.daffa@hotmail.com, ralshahry@stu.kau.edu.sa, ihkhan@kau.edu.sa

ABSTRACT: *Syntactic ambiguity is a common problem in Arabic language. We are exploring the possibility of using corpus-based word collocation data to predict different interpretations of a potentially ambiguous sentence in Arabic. As a case study, we address the problem of disambiguating coordination structures in Arabic to determine how the external modifier (adjective) applies to the coordinated words (nouns) like* القطط والكلاب السوداء *(black cats and dogs). In this paper, we report on an empirical study in which participants were presented with a sequence of trials, each of which consists of potentially ambiguous sentence followed by a comprehension question that relates to the preceding sentence. The study reveals that lexical co-occurrence information, derived using Kilgarriff's Sketch Engine operated on a ca. 1.7 millions words Arabic Web Corpus, can be used to predict the most likely interpretation of a potentially ambiguous sentence.*

Keywords: Arabic word sketches, Arabic web corpus, Ambiguity in Arabic, Coordination ambiguity

Received: 2 June 2015, Revised 21 June 2015, Accepted 1 July 2015

1. Introduction

Ambiguity is a characteristic of a text, whereby the text can be interpreted in more than one different ways. In natural languages (like Arabic or English), ambiguity can arise at different levels. One of the most common type of ambiguity is structural ambiguity (also known as syntactic ambiguity), whereby a sequence of words can be grammatically structured in more than one way hence resulting in more than one interpretation. A typical form of structural ambiguity is *coordination ambiguity*, in which an external modifier occurs in coordinated structures like القطط والكلاب السوداء (black cats and dogs, in English). The phrase القطط والكلاب السوداء is ambiguous, because a reader might be inclined to interpret this as pertaining to both cats and dogs which are black, or only the dogs which are black. Presumably, in this case, majority readers may opt for the former interpretation (i.e. both cats and dogs are black). Now, consider another example الرجال والنساء الملتحي (bearded men and women). In this example, the first interpretation (both men and women are bearded) seems very unlikely.

[1]It is important to mention here that unlike English, in Arabic adjectives succeed the modified noun(s).

Psycholinguistic evidence suggests that in many cases such ambiguities cause confusion (Tanenhaus & Trueswell, 1995), however, sometime, even though theoretically ambiguity can be present in the phrase, most people may interpret the phrase in the same way. Therefore, the problem is how a computer system can determine the likelihood of different interpretations of a given phrase and then decide which interpretation is most likely.

As compared to other forms of structural ambiguity such as prepositional-phrase (PP) attachment ambiguity, coordination ambiguity has received very little attention in Arabic literature. In this paper, we address the problem of disambiguating coordination structures in Arabic to determine how the external modifier applies to the coordinated words or phrases. Arguably, words and phrases of all types can be coordinated. However, to study specific data, we focussed on potentially ambiguous Noun Phrases (NPs) of the form[1] Noun1 and Noun2 Adjective. We call NPs of this form scopally ambiguous, because the scope of Adjective is ambiguous between wide-scope (Adjective applies to both nouns) and narrow-scope (Adjective applies only to Noun2). We estimate which interpretation of a potentially ambiguous NP is most likely for human readers by using statistical information about lexical co-occurrence. The lexical co-occurrence information is derived using Kilgarriff's Sketch Engine (Kilgarriff, Rychly, Smrz, & Tugwell, 2004) operated on ca. 170-million tokens Arabic Web Corpus.

2. Background and Related Work

2.1 Arabic NLP

Arabic natural language processing (ANLP) has gained increasing importance (Daimi, 2001; Othman, Shaalan, & Rafea, 2003; Nwesri, Tahaghoghi, & Scholer, 2005; Shaalan, Rafea, Baraka, & Monem, 2008; Farghaly & Shaalan, 2009; Green & Manning, 2010; Shalaan, 2010; Khan & Siddiqui, 2015), and in recent years a reasonable number of systems have been developed for various applications, including machine translation, information retrieval/extraction, speech synthesis/recognition, text to speech, and tutoring systems. Most ANLP systems developed focus on tools to enable non-Arabic speakers make sense of Arabic texts. For example, tools such as Arabic named entity recognition, machine translation and sentiment analysis are very useful to intelligence and security agencies. Because the need for such tools was urgent, they were primarily developed using machine learning approaches, which usually do not rely on deep linguistic knowledge. Mostly, the NLP tools have been developed for English or other European languages, and because of the specific characteristics of Arabic language such NLP tools are not easily adaptable to Arabic language (Farghaly & Shaalan, 2009). In (Shaalan, Rafea, Baraka, & Monem, 2008), the authors used a grammar-based approach to generate Arabic text, by using interlingua-based spoken dialogue. Their work provides a potential inroad for futuristic research in ANLP.

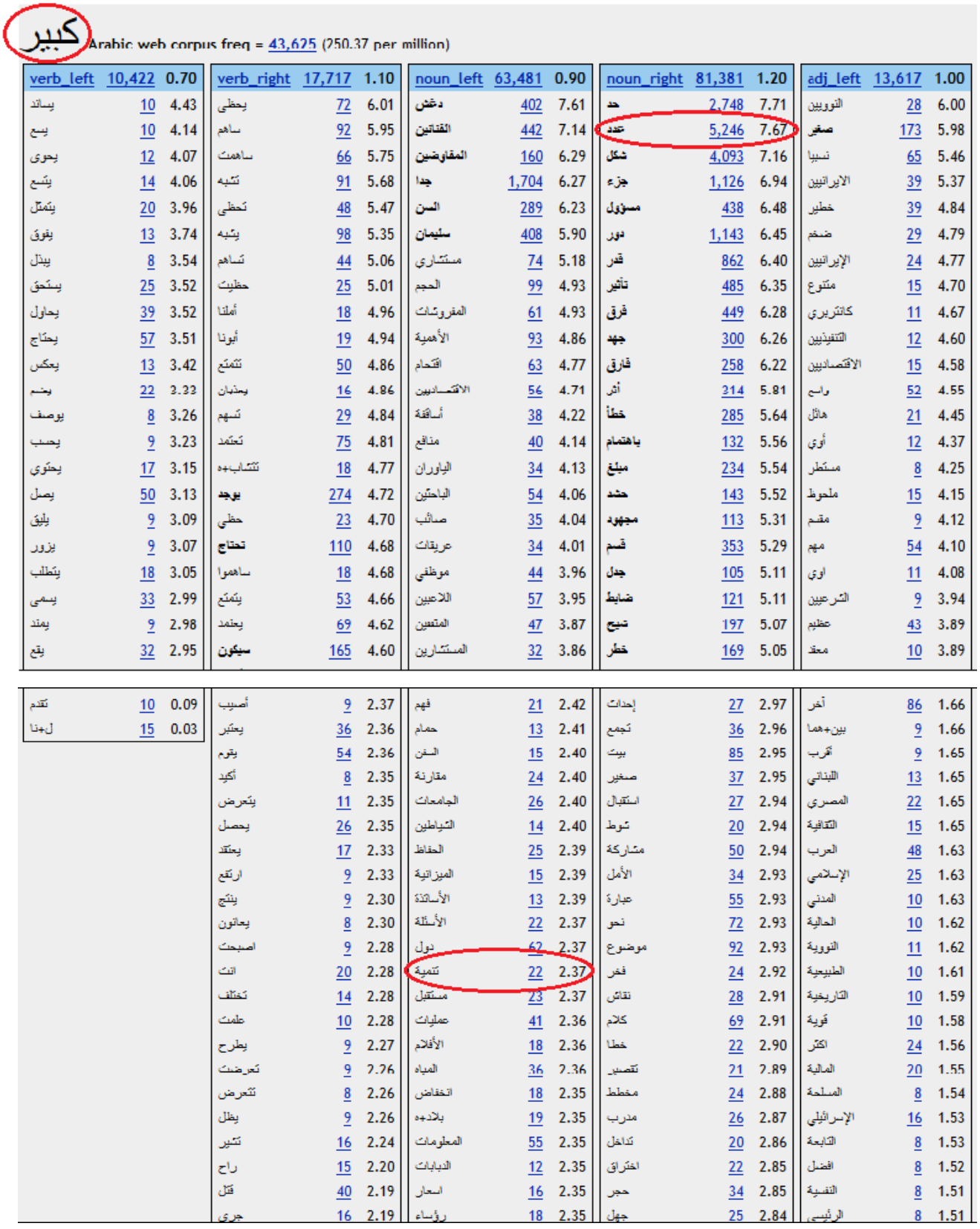
Research in Arabic language reveals that Arabic is a highly inflected language, which constructs its vocabulary through a complicated derivational process using root words (Habash, Introduction to Arabic Natural Language Processing, 2010). These morphological characteristics and various writing styles pose significant challenges in Arabic language analysis tasks (Al-Fares & Roeck, 2000; Rozovskaya, Sproat, & Benmamoun, 2006; Habash, 2006), including ambiguity resolution. For example, the absence of the diacritics could lead to an ambiguous expression, making it extremely difficult to distinguish different words, even in a larger context. The lack of diacritics in most Arabic documents available on the Web is considered as a major challenge to many Arabic NLP tasks.

2.2 Structural Ambiguity in Arabic

The problem of ambiguity in Arabic language has not received serious attention by researchers, mainly due to the special characteristics of Arabic including its high syntactic flexibility (Farghaly & Shaalan, 2009). In literature on ANLP, very few systems have been reported which take ambiguity into account. In a study (Daimi, 2001), the author developed a parser to analyse single-parse Arabic sentences. The parser analyses each sentence and verifies the conditions that govern the existence of certain types of syntactic ambiguities in the sentence. Another interesting piece of work is a chart parser for analyzing Modern Standard Arabic (MSA) sentences (Othman, Shaalan, & Rafea, 2003), which exploits rule-based approach in ANLP (Shalaan, 2010) to satisfy syntactic constraints reducing parsing ambiguity. Grammar rules were developed in which a definite noun object should meet some constraints in order to apply the rules: 1) semantic constraint, the object should be neither a demonstrative noun nor a connected pronoun, and 2) syntactic constraint, the object should be neither a nominative nor genitive case.In yet another study (Green & Manning, 2010), the authors developed a parser informed by a manually annotated grammar for Arabic language. The parser was also evaluated and the results were significantly better than the baseline.

2.3 The Sketch Engine

The Sketch Engine (Kilgarriff, Rychly, Smrz, & Tugwell, 2004) is a corpus query system which can be interfaced with various



كبير Arabic web corpus freq = 43,625 (250.37 per million)

verb_left	10,422	0.70
يساند	10	4.43
يسع	10	4.14
يحوى	12	4.07
يتسع	14	4.06
يتمثل	20	3.96
يفوق	13	3.74
يبذل	8	3.54
يستحق	25	3.52
يحاول	39	3.52
يحتاج	57	3.51
يعكس	13	3.42
يضم	22	3.33
يوصف	8	3.26
يحسب	9	3.23
يحتوي	17	3.15
يصل	50	3.13
يليق	9	3.09
يزور	9	3.07
يتطلب	18	3.05
يسمى	33	2.99
يمتد	9	2.98
يقع	32	2.95
تقدم	10	0.09
ل+نا	15	0.03

verb_right	17,717	1.10
يحظى	72	6.01
ساهم	92	5.95
ساهمت	66	5.75
تتبه	91	5.68
تحظى	48	5.47
يتبه	98	5.35
تساهم	44	5.06
حظيت	25	5.01
أملنا	18	4.96
أبونا	19	4.94
تتمتع	50	4.86
يمتدان	16	4.86
تسهم	29	4.84
تعتمد	75	4.81
تتشابه+ه	18	4.77
يوجد	274	4.72
حظي	23	4.70
تحتاج	110	4.68
ساهموا	18	4.68
يتمتع	53	4.66
يعتمد	69	4.62
سيكون	165	4.60
أصيب	9	2.37
يعتبر	36	2.36
يقوم	54	2.36
أكيد	8	2.35
يتعرض	11	2.35
يحصل	26	2.35
يحتقد	17	2.33
ارتفع	9	2.33
ينتج	9	2.30
يعانون	8	2.30
اصبحت	9	2.28
انت	20	2.28
تختلف	14	2.28
علمت	10	2.28
يطرح	9	2.27
تعرضت	9	2.26
تتعرض	8	2.26
يظل	9	2.26
تثير	16	2.24
راح	15	2.20
قتل	40	2.19
جرى	16	2.19

noun_left	63,481	0.90
دعش	402	7.61
القنانين	442	7.14
المقاومين	160	6.29
جدا	1,704	6.27
السن	289	6.23
سليمان	408	5.90
مستشاري	74	5.18
الحجم	99	4.93
المقروئات	61	4.93
الأهمية	93	4.86
اقتحام	63	4.77
الاقتصاديين	56	4.71
أساقفة	38	4.22
منافع	40	4.14
الياوران	34	4.13
الباحثين	54	4.06
صائب	35	4.04
عريقات	34	4.01
موظفي	44	3.96
اللاعبين	57	3.95
المتعين	47	3.87
المستشارين	32	3.86
فهم	21	2.42
حمام	13	2.41
السفن	15	2.40
مقارنة	24	2.40
الجامعات	26	2.40
الشياطين	14	2.40
الحفاظ	25	2.39
الميزانية	15	2.39
الأساتذة	13	2.39
الأسئلة	22	2.37
دول	62	2.37
تنمية	22	2.37
مستقبل	23	2.37
عمليات	41	2.36
الأفلام	18	2.36
المياه	36	2.36
انخفاض	18	2.35
بلادنا	19	2.35
المعلومات	55	2.35
الدبابات	12	2.35
اسعار	16	2.35
رؤساء	18	2.35

noun_right	81,381	1.20
حد	2,748	7.71
عدد	5,246	7.67
شكل	4,093	7.16
جزء	1,126	6.94
مسؤول	438	6.48
دور	1,143	6.45
قدر	862	6.40
تأثير	485	6.35
فرق	449	6.28
جهد	300	6.26
فارق	258	6.22
أثر	314	5.81
خطأ	285	5.64
باهتمام	132	5.56
مبلغ	234	5.54
حشد	143	5.52
مجهود	113	5.31
قسم	353	5.29
جدل	105	5.11
ضابط	121	5.11
سيح	197	5.07
خطر	169	5.05
إحداث	27	2.97
تجمع	36	2.96
بيت	85	2.95
صغير	37	2.95
استقبال	27	2.94
شوط	20	2.94
مشاركة	50	2.94
الأمل	34	2.93
عبارة	55	2.93
نحو	72	2.93
موضوع	92	2.93
فخر	24	2.92
نقاش	28	2.91
كلام	69	2.91
خطا	22	2.90
تقصير	21	2.89
مخطط	24	2.88
مدرب	26	2.87
تداخل	20	2.86
اختراق	22	2.85
حجر	34	2.85
جهل	25	2.84

adj_left	13,617	1.00
النوويين	28	6.00
صغير	173	5.98
نسبيا	65	5.46
الايرانيين	39	5.37
خطير	39	4.84
ضخم	29	4.79
الإيرانيين	24	4.77
متنوع	15	4.70
كاثتريري	11	4.67
التنفيذيين	12	4.60
الاقتصاديين	15	4.58
واسع	52	4.55
هائل	21	4.45
أوي	12	4.37
مستطر	8	4.25
ملحوظ	15	4.15
مقدم	9	4.12
مهم	54	4.10
اوي	11	4.08
الشرعيين	9	3.94
عظيم	43	3.89
معقد	10	3.89
آخر	86	1.66
بين+هما	9	1.66
أقرب	9	1.65
اللبناني	13	1.65
المصري	22	1.65
الثقافية	15	1.65
العرب	48	1.63
الإسلامي	25	1.63
المدني	10	1.63
الحالية	10	1.62
النووية	11	1.62
الطبيعية	10	1.61
التاريخية	10	1.59
قوية	10	1.58
اكثر	24	1.56
المالية	20	1.55
المسلحة	8	1.54
الإسرائيلي	16	1.53
التابعة	8	1.53
افضل	8	1.52
النفسية	8	1.51
الرئيسي	8	1.51

Figure 1. Word sketches for the adjective (big) كبير

corpora, including different languages for example English and Arabic. One of the key features of the Sketch Engine is its ability to generate summaries of words' grammatical and collocation behaviour.These summaries are called as word sketches. It is instructive to discuss at the outset how word sketches work. The word sketches give information about the frequency with which words are linked by a given grammatical relation. Rather than looking at an arbitrary window of text around a given word, the correct collocations are found by use of grammatical patterns. Suppose we want to generate the word sketches for the Arabic word كبير (big, in English) (node word). Upon receiving this word, the sketch engine provides one list of collocates for each grammatical relation كبير participates in, along with a salience score, which is calculated from the overall frequencies of the node word and the argument word, in the Arabic Web Corpus. The Arabic Web Corpus is part-of-speech tagged corpus, which comprises ca. 170-million tokens (0.4 million words) from modern standard Arabic language. A truncated example is shown in Figure 1. This example shows that, for example, كبير modifies عدد (number) more often (score: 7.67) than خطا (mistake) (score: 2.90); the words of interest are encircled.

3. The Empirical Study

In this study, participants were presented with a sequence of trials, each of which consists a target sentence followed by a comprehension question that relates to the preceding sentence. The target sentences involved an NP with potential scope ambiguity, whereas the comprehension question asks whether the adjective applies to the first noun (i.e. the noun further away from the adjective) or not. For example, for the target sentence رأيت المرأة والرجل الكئيب في السوق (I saw the gloomy man and woman in the market), the comprehension question was هل رأيت المرأة الكئيب في السوق؟ (Did you see the gloomy woman in the market?). A 'yes' answer in this case is considered as a wide-scope reading whereas a 'no' answer implies a narrow-scope reading.

3.1 Material and Design

We define, similar in spirit to Khan, Van Deemter, & Ritchie(2012), that an adjective exhibits high collocation with a noun if the noun appears in the top 20% collocates of the adjective in the adjective-of relationship, produced by sketch engine, operated on the Arabic Web Corpus; a low collocation if the adjective does not appear in the relationship at all. The choice of nouns and adjectives to construct the NPs is motivated by the fact that there is a balanced distribution of adjectives having high (or low) collocation with the two nouns. First, we selected thirty nouns pseudo-randomly to construct sixteen coordinated noun pairs; two nouns were repeated. Then, for each noun pair, word sketches were generated using the sketch difference facility in the sketch engine and an adjective was selected keeping in view the collocational strength of the adjective with the two nouns. A total of 16 adjectives were thus obtained with the following distribution: 4 adjectives having high collocation with both nouns (High-High condition), 4 having low collocation with both nouns (Low-Low condition), 4 having high collocation with the first noun but low with the second (High-Low condition), and 4 having low collocation with the first noun and high with the second noun (Low-High condition). The select adjectives and nouns are shown in Table 1. These adjectives and nouns were then used to construct NPs which were embedded in a one sentence context. This arrangement yielded a total of 16 experimental trials and each participant completed all trials.

3.2 Participants and Procedure

Twenty postgraduate native-Arabic students and employees took part in the experiment voluntarily. A total number of 20 female participants from King Abdul-Aziz University took apart in the experiment; these participants were drawn from different pools: 10postgraduate computer science students, 5 postgraduate economics students, and 5 employees of Arabic language institute with more than 10 years' experience in linguistic correction.

Before running the experiment, the participants were briefed about the purpose and format of the experiment.The trails were presented in a pseudo-random order on a computer screen. First, the experimental sentence appeared on the screen for 15 seconds. During this period, participants read and comprehended the sentence. Then, the sentence disappeared and a comprehension question appeared on the screen; a screenshot of such a sentence is shown in Figure 2. The comprehension questions, which related to the preceding sentence, were designed in a forced choice manner, i.e. participants had to select either a Yes or No answer; a screenshot of such a question is shown in Figure 3.

using survey for locating and identifying syntactic ambiguities has some disadvantages such as not all the students are motivated to fill out survey. As a result, mismatches between real behavior and survey answers could exist.

Condition	Adjective	Noun1	Noun2
High-High	جميلة	فتاة	إمرأة
	فقير	شاب	يتيم
	قصير	فلم	فاصل
	المواقع	تصميم	تطوير
High-Low	شجاع	قائد	طالب
	طويل	نقاش	إستثمار
	الحزب	زعيم	مسؤولي
	الإعلام	وسائل	تصريحات
Low-High	الطازج	الرغيف	البرتقال
	أسود	حمار	حرير
	الوحيد	السبيل	العلاج
	الحياة	مظاهر	جوانب
Low-Low	المستورد	البرتقال	العنب
	المؤمن	حق	مال
	الكئيب	الرجل	المرأة
	الغيب	الإيمان	مفهوم
Noun1 is the nearest noun to the adjective and Noun2 is the noun further away from the adjective; an example NP in the High-Low condition is: طالب شجاعوقائد.			

Table 1. Adjectives and nouns used to construct NPs

4. Results and Discussion

Results were recorded according to whether a participant opted for a wide- or narrow-scope reading. Participants' responses averaged over all four experimental trials per condition are shown in Table 2. The results indicate that when adjective has high collocation with the nearest noun than the noun further away from it (Low-High condition), majority participants (above 78%) opted for a narrow-scope reading. A sign-binomial test further revealed that the difference between wide- and narrow-scope responses are significant ($p < 0.01$). Similarly, when the adjective has high collocation with the second (i.e. further away) noun, participants opted for a wide-scope reading: above 82% in High-High condition and above 73% in High-Low conditions; $p < 0.01$, in both conditions. Interestingly, when adjective exhibits low collocation with both nouns, participants did not opt for a clear preference to one interpretation over the other indicating that such combination of adjectives and nouns could be ambiguous.

The current study revealed some interesting results. Words' collocation data derived from a sizeable corpus can be used to predict the likelihood of different interpretations of potentially ambiguous NPs, in Arabic. On the basis of above results, we conjecture that, in scopally ambiguous NPs, when adjective has high collocation with the noun furthest from the adjective then wide-scope interpretation is more likely irrespective of the collocation strength between adjective and the nearest noun. Similarly, when adjective has high collocation with the nearest noun and low with the furthest noun, then a narrow-scope interpretation is more likely.

Condition	Narrow-scope response	Wide-scope response	P-value
High-High	17.5	82.5	< 0.01
Low-Low	46.38	53.62	= 0.24
Low-High	78.75	21.25	< 0.01
High-Low	26.25	73.75	< 0.01

Table 2. Response proportions (%)

It is worth mentioning here that we used a small and engineered dataset. On the one hand, this allows us to focus on specific and manageable phenomena in a simple experimental design in which every participant is presented with every item. On the other hand, a small dataset can cast doubts on the generalizations which we drew from our sample. However, as the sample NPs were carefully constructed and adjectives were derived from a sizeable Arabic corpus, we are confident that our generalizations are on the right track. Interestingly, our findings corroborate earlier findings ofWillis, Chantree, & De Roeck(2008) and Khan, Van Deemter, & Ritchie(2012)on similar NPs for English language. It is also important to mention here that, in the present study the relationship between the two coordinated nouns is not taken into account. It might be interesting to see if high/low collocation relationship between the two nouns can influence the final interpretation of the NP.

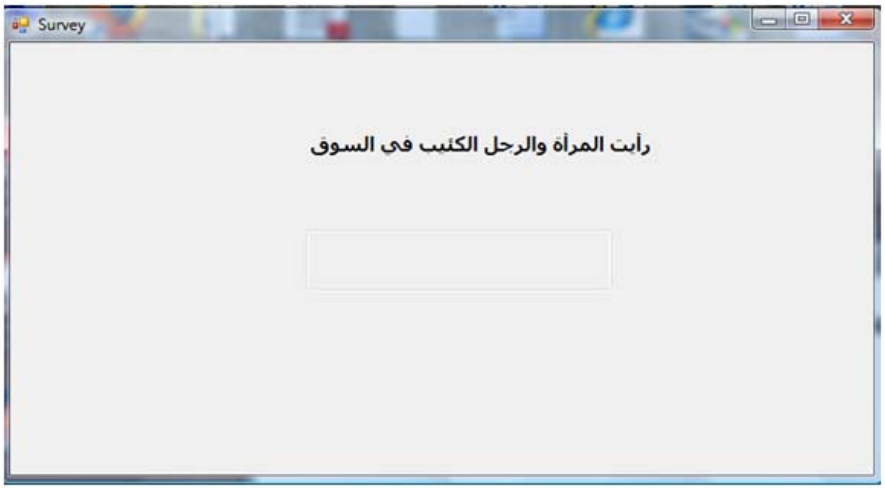


Figure 2. A sample experimental sentence

Figure 3. A sample comprehension question

5. Conclusion

The present study examined empirically the possibility of using corpus-based word collocation data to predict different interpretations of a potentially ambiguous sentence in Arabic. As a case study, we addressed the problem of disambiguating coordination structures in Arabic to determine how the external modifier (adjective) applies to the coordinated words (nouns). The experimental trials were constructed in such a way that there was a balanced distribution of adjectives having high (or low) collocation with the two nouns. The native Arabic participants were presented with a sequence of trials, each of which consists a potentially ambiguous sentence followed by a comprehension question that relates to the preceding sentence. The data revealed that lexical co-occurrence information, derived using Kilgarriff's Sketch Engine operated on Arabic Web Corpus, can be used to predict the most likely interpretation of a potentially ambiguous sentence. More specifically, we conclude that in scopally ambiguous NPs, when adjective has high collocation with the noun furthest from the adjective then wide-scope interpretation is more likely irrespective of the collocation strength between adjective and the nearest nouAn. Similarly, when adjective has high collocation with the nearest noun and low with the furthest noun, then a narrow-scope interpretation is more likely.

Acknowledgement

We thank King Abdulaziz City of Science and Technology (KACST) funding (Grant No. 11-INF- 1520-03) for providing the Sketch Engine licensing. We also thank reviewers for their constructive feedback.

References

[1] Al-Fares, W., Roeck, A. D. (2000). A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots. *ACL '00 Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (p. 199-206). ACM.

[2] Daimi, K. (2001). *Identifying Syntactic Ambiguities in Single-Parse Arabic Sentence.* University of Detroit Mercy: Department of Mathematics and Computer Science.

[3] Farghaly, A., Shaalan, K. (2009). Arabic Natural Language Processing: Challenges and Solutions. *ACM Transactions on Asian Language Information Processing (TALIP)* .

[4] Green, S., Manning, C. D. (2010). Better Arabic Parsing: Baselines, Evaluations, and Analysis. *COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics* (p. 394-402). Stroudsburg, USA: ACM.

[5] Habash, N. (2006). Arabic Tutorial. *The Fifth International Conference on Language Resources and Evaluation, LREC'06, 2006*.

[6] Habash, N. (2010). *Introduction to Arabic Natural Language Processing.* Morgan & Claypool Publishers.

[7] Khan, I. H., Siddiqui, M. A. (2015). Do speakers produce different referring expressions in their native language than a non-native language? . *International Journal of Computational Language Research* , To appear.

[8] Khan, I. H., Van Deemter, K., Ritchie, G. (2012). Managing Ambiguity in Reference Generation: The Role of Surface Structure. *Cognitive Science Society* .

[9] Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. *Proceedings of EURALEX*. CiteSeerX.

[10] Nwesri, A., Tahaghoghi, S., Scholer, F. (2005). Stemming Arabic Conjunctions and Prepositions. *12th International Conference on String Processing and Information Retireval* (p. 206–217). Buenos Aires, Argentina: Springer.

[11] Othman, E., Shaalan, K., & Rafea, A. (2003). A Chart Parser for Analyzing Modern Standard Arabic Sentence. *MT Summit IX Workshop on Machine Translation for Semitic Languages: Issues and Approaches* (p. 37–44). New Orleans, Louisiana, USA: ACL.

[12] Rozovskaya, A., Sproat, R., Benmamoun, E. (2006). Challenges in Processing Colloquial Arabic: The challenge of Arabic for NLP/MT. *In international Conference at the British Computer Society* (p. 4 - 14). London: CiteSeer.

[13] Shaalan, K., Rafea, A., Baraka, H., Monem, A. A. (2008). Generating Arabic Text from Interlingua. *Proceedings of the 2nd workshop on computational approaches to Arabic script-based languages* (p. 137-144). Stanford, USA: Linguistic Society of America.

[14] Shalaan, K. (2010). Rule-based Approach in Arabic Natural Language Processing. *International Journal on Information and Communication Technologies* .

[15] Tanenhaus, M., & Trueswell, J. (1995). Sentence Comprehension. In B. M. Bly, & D. E. Rumelhart, *Handbook of Perception and Cognition* (pp. 217-262). NY: Academic Press.

[16] Willis, A., Chantree, F., & De Roeck, A. (2008). Automatic Identification of Nocuous Ambiguity. *Springer Netherlands* .