Evaluation of automatic collocation extraction methods for language learning

Vishal Bhalla
Department of Informatics
Technical University of Munich
Arcisstraße 21, 80333
München, Germany
vishal.bhalla@tum.de

Klara Klimcikova
Department of English and American Studies
Ludwig Maximilian University of Munich
Schellingstrae 3, 80799
München, Germany
k.klimcikova@lmu.de

Abstract

A number of methods have been proposed to automatically extract collocations, i.e., conventionalized lexical combinations, from text corpora. However, the attempts to evaluate and compare them with a specific application in mind lag behind. This paper compares three end-to-end resources for collocation learning, all of which used the same corpus but different methods. Adopting a gold-standard evaluation method, the results show that the method of dependency parsing outperforms regex-over-pos in collocation identification. The lexical association measures (AMs) used for collocation ranking perform about the same overall but differently for individual collocation types. Further analysis has also revealed that there are considerable differences between other commonly used AMs.

1 Introduction

Collocations, as the most common manifestation of formulaic language, have attracted a great deal of research in the last decade (Wray, 2012). Most of the research on collocations has been connected to their definition (section 2.1) and extraction (section 2.2), but also to their acquisition, and consequently teaching. Herbst and Schmid (2014) argue, "Any reflection upon what is important in the learning and, consequently, also in the teaching of a foreign language will have to take into account the crucial role of conventionalized but unpredictable collocations. Any attempt by a learner to achieve some kind of near-nativeness will have to include facts of language such as the fact that it is *lay* or *set the table* in English, but *Tisch decken* in German, and *mettre la table* in French" (p. 1).

Collocation learning comes down to three main benefits for language learners: accurate production, efficient comprehension and increased fluency of processing (e.g., Men, 2017; Durrant and Mathews-Aydınlı, 2011). To increase their language proficiency, beginners and advanced learners often look up for words and it's common collocates online, using a mobile app or web browser and it benefits to provide personalized items, tailored to the user's interest and proficiency. Examples of using collocations for building educational applications include question generation (e.g., Lin et al., 2007), distractor generation (e.g., Liu et al., 2005; Lee and Seneff, 2007) for multiple choice cloze items and an online collocation writing assistant - Collocation Inspector (Wu et al., 2010a) in the form of a web service.

Despite their widely recognized importance and ubiquity in language use, collocations pose a great challenge for language learners thanks to their arbitrary nature and the learner's insufficient experience with the target language (Ellis, 2012). Thus, there is a pressing need to create resources for language learners to support their explicit collocation learning. Given the vast amount of collocations and the different goals of language learners, various methods have been proposed to extract them automatically from text. Yet it is still not conclusive which one performs the best for language learning and "the selection of one or another seems to be somewhat arbitrary" (González Fernández and Schmitt, 2015) (p. 96).

This paper[1] attempts to evaluate three end-to-end resources of collocations built for language learning: Sketch Engine[2] (Kilgarriff et al., 2014), Flexible Language Acquisition (FLAX)[3] (Wu, 2010) and Elia (Bhalla et al., 2018). They use the same British Academic Written English (BAWE) corpus (Nesi, 2011), but different meth-

[1] Code, data and evaluation results are available at https://github.com/vishalbhalla/autocoleval

[2] https://www.sketchengine.eu/

[3] http://flax.nzdl.org/greenstone3/flax

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 264–274
Florence, Italy, August 2, 2019.

ods for collocation identification, i.e., regex-over-pos, n-grams combined with regex-over-pos and dependency parsing, respectively, and also different association measures for collocation ranking, i.e., Log Dice, raw frequency and Formula Teaching Worth (FTW). On top of that, we compare other widely used lexical association measures of MI, MI2, MI3, t-score, log-likelihood, Salience and Delta P using the data from the best performing candidate identification method as a baseline. For our evaluation, we use the expert-judged Academic Collocation List (ACL) (Ackermann and Chen, 2013) as a reference set (section 3.1), and calculate the recall and precision metrics separately for collocation identification and ranking.

2 Theoretical Background

2.1 Notion of Collocation

Among the many different interpretations of collocations in the literature, three leading approaches can be distinguished: psychological, phraseological and distributional (Men, 2017).

The psychological approach envisages collocations as lexical associations in the mental lexicon of language users underlying their fluent and meaningful language use (e.g., Ellis et al., 2008). This perspective on collocations is supported by the evidence from psycholinguistic research using reaction time tasks, free associations tasks, self-paced reading and eye-tracking which suggests that collocations are holistically stored as chunks and thus processed faster (Wray, 2012). However, as found out by Meara (2009), the storage of word associations in the mental lexicon of native speakers is different from that of nonnative speakers.

The phraseological approach focuses predominantly on delimiting collocations (*call a meeting*) from free word combinations with a predictable meaning (*call a doctor*), on the one hand, and fixed idioms with an unpredictable meaning (*call it a day*) on the other (e.g., Cowie, 1998) by defining a set of criteria related to the compositionality of meaning and fixedness of form. Schmitt (2010) argues that such approach is rather problematic for the identification task as it is not clear how to operationalize such criteria without making it subjective and labor-intensive.

The distributional approach, also called Firthian or frequency-based, shifts the focus from the semantic aspects of collocations to structural. As Sinclair (1991) put it, "Collocation is the co-occurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening" (p. 170). Following this definition, various criteria have been considered for identifying collocation, e.g., distance, frequency, exclusivity, directionality, dispersion, type-token distribution and connectivity (Brezina et al., 2015). However, some researchers (e.g., Bartsch, 2004) argue that because of the little account of syntactic features of the words, it fails to capture certain collocations, e.g., the collocation *collect stamps* in the sentence *They collect many things, but chiefly stamps*, or vice versa, captures false collocations, such as *things but*.

Despite the obvious differences, there is considerable overlap between the three approaches as Durrant and Mathews-Aydnl (2011) rightly point out, "Non-compositionality and high frequency of occurrence can both be cited as evidence for holistic mental storage, and non-substitutability of parts can be evidenced in terms of co-occurrence frequencies in a corpus" (p. 59). It is precisely this extended notion of two-word collocations which was adopted by the collocation references under investigation in this study.

2.2 Automatic Extraction of Collocations

The task of collocation extraction is usually split into two steps, that of candidate identification which automatically generates a list of potential collocations from a text according to some criteria, and that of candidate ranking, which ranks the list to keep the best collocations on top according to some association measure (Seretan, 2008).

2.2.1 Candidate Identification

In the candidate identification step, four prominent methods can be distinguished based on the proximity of words and the amount of linguistic information used: window, n-gram, regex-over-pos and parsing. The first two are based on linear proximity whereas the other two are on syntactic proximity.

The window-based method (e.g., Brezina et al., 2015) identifies collocations within a window of n words before and after the target word. It belongs to the most commonly known and used and directly follows the Firthian definition of collocations. Similarly, the n-gram method (e.g., Smadja and McKeown, 1990), extracts sequences of adjacent n words including the target word. The appli-

cation of these two methods can vary along several dimensions, e.g., the nature of words considered, such as word forms, lemmas or word families (Seretan, 2008), the context span on the left and right or the number of grams, part-of-speech filtering, etc. However, due to the lack of linguistic information used, these methods are prone to many recall and precision errors (for a detailed discussion, see Lehmann and Schneider, 2009).

In contrast to the previous two methods, the regex-over-pos method takes into account the grammatical relations between words (e.g., Wu, 2010). It identifies collocations in text via regular expressions over part-of-speech tags which match a certain grammatical pattern of the collocation. An alternative, though less frequent, method identifies collocations in a syntactic relation via parsing (Seretan, 2008), and thus accounts for the syntactic flexibility feature of collocations. Bartsch and Evert (2014) found out that collocation extraction using parsing method improved the results in comparison to the window method. However, they also caution that the success depends on the accuracy of the parser and the set of grammatical relations used.

2.2.2 Candidate Ranking

The next step of candidate ranking entails measuring the strength of association between the two words, hence association measure (AMs). In principle, AMs compare the observed and expected frequencies of collocations in different ways, and thus differ in how much they highlight or downplay different features of collocations (for a detailed overview, see Pecina, 2010). There is no single best performing AM but rather the choice of an appropriate measure depends on the particular purpose and theoretical criteria. In language learning research and practice, the following AMs have received most attention: raw frequency, MI, MI2, MI3, Log Dice, t-score, log-likelihood, Salience, FTW and Delta P.

The Mutual Information (MI) measure prioritizes rare exclusivity of collocations which is strongly linked to predictability (Gablasova et al., 2017). However, it is also biased towards low-frequency combinations which can be circumvented by setting a minimum frequency threshold or giving extra weight to the collocation frequency by squaring (MI2) or cubing (MI3).

The Log Dice score is similar to MI2 and highlights the exclusivity of word combinations without putting too much weight to rare combinations. However, Log Dice, in contrast to MI2, is suitable for comparing scores from different corpora and has been described as a "lexicographer-friendly association score" (Rychlỳ, 2008, p. 6-9). Another measure adjusted for lexicographic purposes is Salience, the forerunner of Log Dice, which combines the strengths of MI and log frequency (Kilgarriff and Tugwell, 2002).

The t-score represents the strength of association between words by calculating the probability that a certain collocation will occur without considering the level of significance (Pecina, 2010). It prioritizes the frequency of the whole collocation, and hence there is a tendency for frequent collocations to rank higher.

The only measure created specifically for pedagogical purposes is the Formula Teaching Worth (FTW) which is again a combined measure of MI and the raw frequency with more weight given to the former. It was derived from an empirical research using both statistical measures and instructor judgments. Basically, the score represents "a prediction of how instructors would judge their teaching worth" (Simpson-Vlach and Ellis, 2010, p. 496).

In contrast to the previous measures, log-likelihood is a statistic which determines whether the word combination occurs more frequently than chance or not. In particular, the score does not provide information on "how large the difference is" but rather "whether we have enough evidence in the data to reject the null hypothesis" (Brezina et al., 2015, p. 161).

The last measure is Delta P which takes directionality into account and calculates the strength of the attraction between two words for each word separately. Therefore, in contrast to all the previous measures, it does not treat the collocational relationship as symmetrical (Gries, 2013).

3 Methodology

3.1 Reference Set

The recently compiled Academic Collocation List (ACL) (Ackermann and Chen, 2013) was selected as the reference set (gold standard) to be compared against the test sets. Five main considerations drove this decision: First, it needed to be in line with the nature of the BAWE[4] corpus that was cho-

[4] https://www.coventry.ac.uk/research/research-directories/

sen as a source input for extracting collocations. BAWE contains around 3000 good-standard student assignments (with 6,506,995 words), evenly distributed across four broad disciplinary areas (Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences) and across four levels of study (undergraduate and taught masters level). Since BAWE is a collection of academic writing of university students, the baseline set should also consist of academic collocations. Second, it should contain collocations consisting of two words as all the three resources focus on two-word collocations. Third, the collocations should preferably be grouped into collocation types based on their word classes or syntactic functions as in the test sets. Fourth, the reference set should be human-made or human-judged to ensure the quality of collocations. And finally, it should be compiled for pedagogical purposes.

The ACL comprises of 2,469 lexical collocations in written academic English and is based on a written part of the Pearson International Corpus of Academic English (PICAE) of around 25 million words. It was carefully compiled using the combination of automatic computational analysis to ensure an adequate recall and human judgment to ensure the quality and relevance of the collocations for pedagogical purposes. Consisting of the most frequent and pedagogically relevant entries, ACL can therefore be immediately operationalized by English for Academic Purposes (EAP) teachers and students. By highlighting the most important cross-disciplinary collocations, the ACL can help learners increase their collocational competence and thus their proficiency in academic English. The collocations are grouped into eight collocation types: adjective + noun, noun + noun, verb + noun, verb + adjective, adverb + verb, verb + adverb, adverb + verb past participle, adverb + adjective.

To make it comparable to the test sets, we lemmatized all its inflected word forms using an automatic lemmatization tool from SpaCy[5] and then manually checked all the errors. Next, it was organized by headwords with the POS tags *noun*, *adjective* and *verb*, grouped by possible collocation types, and with the respective collocates appended resulting in a list of 1,455 headwords, 11 collocation types and 4,626 collocations as presented in Table 1. For example, the notation of the collocation type *n2_adj1* indicates that the headword is a noun (*n*) in the *2nd* position in the collocation pair, and the collocate is an adjective (*adj*) in the *1st* position, so when the learner searches the adjectival collocates for the word *feature*, it gives him the collocate *distinguishing* among others.

Collocation Type	Headwords	Collocations
n1_n2	39	62
n2_n1	52	62
n2_v1	156	306
n2_adj1	483	1769
v1_n2	107	306
v1_adj2	8	30
v1_adv2	19	29
v2_adv1	79	139
adj1_n2	416	1769
adj2_v1	23	30
adj2_adv1	73	124
Total	1455	4626

Table 1: Reference set grouped by collocation types starting with a headword where noun is *n*, adjective is *adj*, verb is *v*, adverb is *adv* and the numbers *1* and *2* indicate their positions in the collocation pair.

3.2 Tests sets

3.2.1 Sketch Engine

Sketch Engine (SE) is an online corpus software with a wide range of functions and preloaded corpora which can be used for pedagogical purposes either indirectly, in the creation of textbooks and dictionaries, or directly in the classroom (Kilgarriff et al., 2014). One of its functions is the Word Sketch for extracting collocations in a range of grammatical patterns, and one of its corpora is BAWE. The corpus is automatically POS-tagged using CLAWS 7[6] and the collocations are identified with the help of their embedded Sketch Grammar[7] which is a set of regular expressions over POS tags. The retrieved collocates are then organized based on the grammatical relation to the headword and within each relation sorted by the Log Dice measure (alternatively, raw frequency).

The SE collocations were extracted using web scraping wherein, firstly, the URL was built us-

current-projects/2015/british-academic-written-english-corpus-bawe/

[5] https://spacy.io/

[6] https://www.sketchengine.eu/english-claws7-part-of-speech-tagset/

[7] For a full list of Sketch Grammar, see https://the.sketchengine.co.uk/corpus/wsdef?corpname=preloaded/bawe2

ACL	Sketch Engine	FLAX	Elia
n1_n2	modifies	n-nn	NOUN + NOUN
n2_n1	modifier	n-nn	NOUN + NOUN
n2_v1	object_of	n-vn	VERB + NOUN
n2_adj1	modifier	n-an	ADJ + NOUN
v1_n2	object	v-vn	VERB + NOUN
v1_adj2	adj_comp np_adj_comp	v-vppa	VERB + ADJ
v1_adv2	modifier	v-vr	VERB + ADV
v2_adv1	modifier	v-rv	VERB + ADV
adj1_n2	modifies	a-an	ADJ + NOUN
adj2_v1	adj_comp_of np_adj_comp_of	a-vppa	VERB + ADJ
adj2_adv1	modifier	a-ra	ADV + ADJ

Table 2: Mapping of collocation types between the reference set (ACL) and test sets (Sketch Engine, FLAX, Elia)

ing the lemma and POS tag of each word and then the eleven collocation types from the reference set were mapped to the collocation types used at SE to pickup the collocations of interest (Table 2). Picking up all the headwords (lemmas) from the reference set, the count and score of each collocate was stored in an intermediate file for each lemma in order to generate SE files[8] for the final evaluation.

3.2.2 FLAX

FLAX (Flexible Language Acquisition) is an online library and tool specifically created for collocation learning (Wu, 2010). It consists of large collections of collocations and phrases extracted from different corpora, one of which is BAWE, and can be used for searching collocations for a particular word or for automatic generation of a variety of collocation exercises and games. The collocations are extracted using the combination of n-gram and regex over-pos methods which involved the following steps. Firstly, n-grams (n=5) are extracted from the corpus and tagged with the OpenNLP[9] tagger. The tagged 5-grams are then matched against a set of regular expressions based on predefined collocation types[10]. Finally, the individual collocations organized by collocation types are then sorted by raw frequency within each collocation type (Wu, 2010, p. 98).

For the evaluation, all FLAX collocations[11] were extracted using the same web scraping process as for SE. However, as FLAX operates on word forms in contrast to the reference set operating on lemmas, all lemmas from the reference set had to be converted to their word forms using Pattern[12] (Smedt and Daelemans, 2012) to get the corresponding collocations from FLAX for each headword in the reference set and then remapped back to its lemma to continue with the same evaluation flow as in SE.

3.2.3 Elia

Elia[13] is an intelligent personal assistant for language learning which provides immediate assistance for English learners when they use English online (Bhalla et al., 2018). One of its design features is to provide a learner with a list of collocates for a given word, which are in line with the learner's proficiency level. It is based on BAWE where, firstly, all the dependency relations using the SpaCy parser[14] are extracted and mapped to a predefined set of 15 collocation types and then run for the Academic Vocabulary List (Gardner and Davies, 2013) of 20,000 most frequent academic words from the Corpus of Contemporary American English (COCA)[15]. Subsequently, the collocations are organized for each headword (lemma)

[8] Code and data in the 'sketchengine' folder of the Supplementary Material.

[9] http://opennlp.apache.org/

[10] For a full list of the collocation types with examples, see Wu et al. (2010b, p. 9).

[11] Code and data in the 'flax' folder of the Supplementary Material.

[12] https://github.com/clips/pattern

[13] Elia is not available online yet, however, the code to generate the database of collocations can be accessed at https://drive.google.com/open?id=1FGFy_yp797saphx8-wzcLkMxQbkCVZlp

[14] https://spacy.io/usage/linguistic-features#section-dependency-parse

[15] https://corpus.byu.edu/coca/

and collocation type and ranked according to the Teaching Worth Formula (Simpson-Vlach and Ellis, 2010).

For the evaluation, only the collocations for the headwords and collocation type present in the reference set were filtered out from Elia after running the code from the link shared previously. On running this setup, intermediate files for each headword containing all its collocates along with the chosen metric were generated. These are in line with the web scraping files from Sketch Engine and FLAX in order to generate the final evaluation files[16].

4 Results and Discussion

For the comparative evaluation of the three test sets, the standard metric recall and precision were calculated separately for identification and ranking of collocations grouped into collocation types. On top of that, additional evaluation was performed on the best performing test set as a baseline to compare different collocation ranking measures introduced in section 2.2.2.

4.1 Candidate Identification

Table 3 clearly shows that the method of dependency parsing used by Elia resulted in higher overall recall (99%) than the method of regex-over-pos used by Sketch Engine (91%) and FLAX (84%). It seems that some dependency parsers have reached a sufficiently high accuracy to be used for collocation extraction or other NLP tasks (Levy et al., 2015). At the same time, there are obvious differences between Sketch Engine and FLAX, despite using the same method (regex-over-pos), which leads to the conclusion that manual mappings of collocation types and syntactic patterns might be as important as the method itself. Another plausible explanation could be the fact that FLAX used regex patterns over 5-grams extracted from the corpus whereas Sketch Engine over full sentences.

Turning to individual collocation types (CTs), all of them achieved a high recall of above 80% in all three test sets, except for *v1_adj2* and *adj2_v1* in FLAX with a recall of only 13% and 7% respectively. Tempting as it might seem, this does not explain the lowest overall recall for FLAX as they account for only 7% (54 out of 710) of all missed collocations. FLAX performed especially well for *v1_adv2* (100%) in comparison to its other CTs starting from 90% *(v2_adv1)* downwards to 7% *(adj2_v1)*. On the other hand, the results for Sketch Engine are rather consistent across individual CTs ranging from 87% *(adj2_adv1)* to 94% *(n2_v1)*. The same applies for Elia ranging from 98% *(n1_n2, n2_n1)* to 100% *(v1_adj2, adj2_v1, v1_adv2)*.

	SE		FL		EL	
Collocation Type	R	P	R	P	R	P
n1_n2	89	6	82	2	98	1
n2_n1	89	5	81	1	98	0
n2_v1	94	8	88	3	99	1
n2_adj1	88	10	86	4	99	2
v1_n2	92	4	84	2	99	1
v1_adj2	90	10	13	2	100	1
v1_adv2	90	5	100	4	100	1
v2_adv1	92	9	90	5	99	2
adj1_n2	93	6	84	5	99	2
adj2_v1	90	40	7	9	100	7
adj2_adv1	87	10	89	8	99	5
Total	91	**7**	84	4	**99**	2

Table 3: Candidate identification comparison of Sketch Engine (SE), FLAX (FL) and Elia (EL) across collocation types with the recall (R) and precision (P) values in percentages.

Looking closer at the results for Elia, we found out that exactly one half (19) of all the missed collocations (38) was due to parsing or tagging errors whereas the other half was due to different type classification; for example, the collocation *learning activity* was grouped under *n2_adj1* in the reference set whereas, in Elia, it was assigned to *n1_n2*, and thus missed. This might as well be the case for some of the missed collocations in Sketch Engine and FLAX.

The precision, on the other hand, is very low for all (the highest 7% reached by Sketch Engine) at the expense of high recall. This, however, is not that important at this stage since the next step of ranking should shift all the irrelevant collocations to the bottom.

4.2 Candidate Ranking

For candidate ranking, recall and precision values were calculated for three samples of *n*-best candidates per headword for each test set: Top 4,626 where *n* refers to the exact number of collocates

[16]Code and data in the 'elia' folder of the Supplementary Material.

	Top 4,626			Top 14,550						Top 29,100					
	SE	FL	EL	SE		FL		EL		SE		FL		EL	
Collocation Type	R=P	R=P	R=P	R	P	R	P	R	P	R	P	R	P	R	P
n1_n2	19	15	10	58	10	52	8	37	6	77	8	73	6	68	5
n2_n1	31	23	19	69	10	61	7	56	7	82	8	68	4	69	4
n2_v1	42	44	48	73	15	70	14	83	16	87	10	75	8	90	9
n2_adj1	36	41	39	50	18	52	19	50	18	68	13	66	12	65	12
v1_n2	37	31	35	55	16	47	14	52	15	68	10	57	8	64	9
v1_adj2	47	0	43	73	28	3	4	63	24	80	17	7	5	73	14
v1_adv2	48	34	34	86	13	86	13	83	13	90	8	97	8	86	7
v2_adv1	46	52	37	81	15	79	15	80	14	91	11	86	9	90	8
adj1_n2	36	43	44	40	18	47	21	49	20	57	13	61	14	64	14
adj2_v1	70	3	33	90	29	7	9	97	15	90	29	7	9	100	10
adj2_adv1	37	54	28	80	16	82	17	80	14	87	11	85	11	89	8
Total	37	**41**	40	51	**17**	52	**17**	**54**	**17**	67	**12**	65	11	**68**	11

Table 4: Candidate ranking comparison of Sketch Engine (SE), FLAX (FL) and Elia (EL) across collocation types for three samples: Top 4,626 (*n*-best collocates per headword where *n* refers to the number of collocations per headword in the reference set), Top 14,550 (10-best collocates per headword) and Top 29,100 (20-best collocates per headword) with the recall (R) and precision (P) values in percentages. Note that the recall and precision results for the top 4,626 are the same (i.e. R=P) because the number of missed collocations (false negatives) and unwanted collocations (false positives) is the same. And this is because the number of the TOP collocations in the first test sample (4,626) is the same as the total number of collocations in the reference set (4,626).

per each headword in the reference set, Top 14,550 to the 10-best collocates per headword, and Top 29,100 to the 20-best collocates per headword.

As illustrated in Table 4, the association measure Log Dice used by Sketch Engine performed slightly worse (37%) overall than Elia (40%) using FTW, a combination of MI and frequency, and FLAX (41%) using raw frequency for the Top 4,626 sample. As the sample increased to 14,550, Elia with a recall of 54% outperformed FLAX (52%) and Sketch Engine (51%). In the even larger sample of 29,100, Elia was still marginally better reaching 68% whereas Sketch Engine outperformed FLAX with a recall of 67% and 65% respectively. It seems that Log Dice improves its performance as more of the data is examined whereas raw frequency acts in quite the opposite way. However, it should also be pointed out that the differences between all of the scores are very subtle, less than 4% in all the samples. This is even more pronounced in the overall precision results which, for all three resources, are the same (17%) in Top 14,550 and almost the same (12%, 11%, 11%) in Top 29,100.

Looking at the individual CTs, an interesting picture of differences emerges. Sketch Engine's measure performed consistently better for *n1_n2*, *n2_n1*, *v1_n2* and *v1_adj2* in all three samples. Elia's measure performs consistently better for *n2_v1* and *adj1_n2*. FLAX seems to perform better only for *v2_adv1* and *adj2_adv1* for Top 4,626 but it is not consistent for the other samples. Variability can be found not only among individual resources but also among individual CTs within one resource. For example, in Top 4,626, Sketch Engine reaches a recall of 19% for *n1_n2* and of as high as 70% for *adj2_v1*. Recall values for Elia range from 37% (*n1_n2*) to 97% (*adj2_v1*) and for FLAX from 7% (*adj2_v1*) to 86% (*v1_adv2*) in Top 14,550. The syntactic structure underlying collocations seems to have a great impact on the results, and thus should always be considered and specified as already suggested in some previous studies (e.g., Evert and Krenn, 2001; Bartsch and Evert, 2014).

To sum it up, despite the apparent similarities in the overall recall and precision values, it would be misleading to conclude that the three measures are equally efficient since they had a different data from the identification step to start with. It becomes clear when looking at the individual collocation types, for example *v1_adj1* where Elia

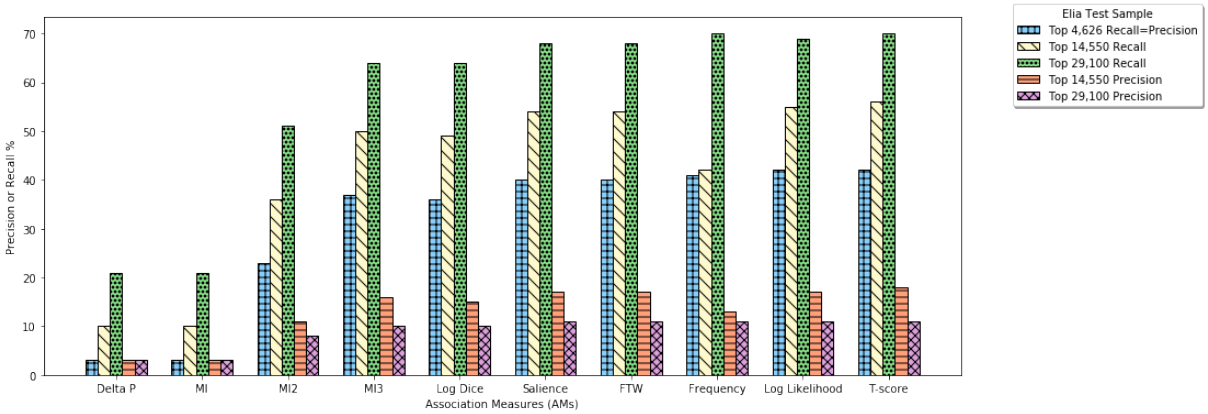

Figure 1: Comparison of different association measures (AMs) using Elia as a baseline.

reached 43% as compared to 0% by FLAX. This could have been caused by the low recall (13%) of FLAX in the identification part. The issue with credit assignment is that it is not clear how much of the success can be attributed to the identification method discussed in the previous section and how much to the metric itself. To exclude the identification method as a factor, we decided to perform another analysis: the comparison of different AMs using the best-performing data from the candidate identification step, that is Elia, as a baseline to find out the differences when all things being equal.

4.2.1 Comparison of Different AMs with Elia as a Baseline

Using Elia collocations as a baseline, we have computed recall and precision for ten different ranking measures described in section 2.2.2. As in the previous section, we have computed it separately for the three samples of Top 4,626, Top 14,550 and Top 29,100, however, this time not for all individual CTs separately. The formulas used to compute each of the different AMs can be found in Brezina et al. (2015, p. 169-170).

The results on candidate ranking, arranged progressively by the best measures in Figure 1, show that recall curves for all AMs increase whereas precision curves decrease with the increased sample sizes as expected. In terms of coverage, the best performing measures are t-score and log likelihood across all samples with the recall values of 42%, 56%, 70% and 42%, 55%, 69% respectively. They are followed by Salience and FTW with the same values of 40%, 54%, 68%. All of these four measures exhibit a consistent behavior increasing by about 14% with the increased samples. On the other hand, the raw frequency measure, even though reaching similarly high scores for Top 4,262 (41%) and Top 29,100 (70%), increases only by 1% for Top 14,550. The next two measures MI3 and Log Dice lag slightly behind with the scores of 36%, 50%, 64% and 36%, 49%, 64% respectively, consistently increasing by about 14%. The MI2 score performs significantly worse with a recall of 23%, 36%, 51%. The most collocations are missed by the measures Delta P and MI both reaching only a recall of 3%, 10%, 21%.

The precision values defining quality of the collocations point to very similar tendencies with t-score and log likelihood reaching the highest precision in all three samples with the scores of 42%, 18%, 11% and 42%, 17%, 11% respectively and with the FTW and Salience measures right behind, both with 40%, 17%, 11%. MI3 and Log Dice performs about the same with 37%, 16%, 10% and 36%, 15%, 10% respectively. Again, the MI2 score misses significantly more collocations than the previous measures reaching a precision of 23%, 11%, 8%. Surprisingly enough, MI and Delta P, both reached the lowest precision score of 3% for all samples. Thus, it can be concluded that the sample size does not affect the precision of the MI and Delta P association measures which, in the case of MI, is consistent with the previous findings by Evert and Krenn (2001).

These results question the dominant role of MI for collocation extraction (Gablasova et al., 2017), at least for language learning purposes. It also questions the assumption that Log Dice is fairly similar to MI or MI2 as our results suggests that it is actually more similar to MI3 (Gablasova et al., 2017). Furthermore, Delta P did not fulfill the expectations as expressed by Gries (2013). However, in defense of Delta P, it must be pointed out

that the reference list did not indicate the direction of attraction for collocations, which is the underlying assumption of the Delta P measure, which might be the reason for the poor results. On the other hand, there are only subtle differences between some of the best-performing measures, such as log likelihood and t-score or FTW and Salience.

5 Conclusion

The aim of this study was to evaluate three collocation learning resources namely Sketch Engine, FLAX and Elia on a pedagogical reference - Academic Collocational List, where all of them use the same corpus of academic writings of university students but different methods for collocation identification and different lexical association measures for collocation ranking.

The findings indicate that using dependency parsing (Elia) for collocation identification led to much better results than using regular expressions over tagged corpus (Sketch Engine and FLAX). However, the success does not depend on the specific method entirely, but also on the quality of the set of syntactic structures. Using the same method with differently designed collocation types might lead to very different results, as was the case for Sketch Engine and FLAX.

The evaluation of collocation ranking has revealed that, overall, some of the association measures perform equally well, such as t-score, log-likelihood, FTW (used by Elia) and Salience. Raw frequency (used by FLAX) was also found to perform well but acting inconsistently across different sample sizes. The Log Dice measure (used by Sketch Engine) worked best for the majority of individual collocation types in comparison to raw frequency and FTW. On the other hand, the widely used MI and newly introduced Delta P were relatively poor in comparison to other AMs, but exhibited consistency in precision across varying sample sizes.

It has also become apparent that there are considerable differences between individual collocation types, and therefore should always be considered as a factor in collocation extraction. However, a future line of work is required to substantiate the consistency of these results on different reference lists and corpora.

Acknowledgments

We would like to thank Mrs. Wu for her insights on FLAX extraction, Aisulu for preparing the COCA list, Ivet for the help in large scale experiments and all the anonymous reviewers for their critical feedback.

References

Kirsten Ackermann and Yu-Hua Chen. 2013. Developing the Academic Collocation List (ACL)–A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12(4):235–247.

Sabine Bartsch. 2004. *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunter Narr Verlag.

Sabine Bartsch and Stefan Evert. 2014. Towards a Firthian notion of collocation. *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2nd Work Report of the Academic Network Internet Lexicography, OPAL–Online publizierte Arbeiten zur Linguistik. Institut für Deutsche Sprache, Mannheim, to appear*.

Vishal Bhalla, Klara Klimcikova, and Aisulu Rakhmetullina. 2018. The missing link between learners' language use and their language learning - Elia. In *Proceedings of the 13th Teaching and Language Corpora Conference*.

Vaclav Brezina, Tony McEnery, and Stephen Wattam. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2):139–173.

Anthony Paul Cowie. 1998. *Phraseology: Theory, analysis, and applications*. OUP Oxford.

Philip Durrant and Julie Mathews-Aydınlı. 2011. A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, 30(1):58–72.

Nick C Ellis. 2012. Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, 32:17–44.

Nick C Ellis, RITA Simpson-Vlach, and Carson Maynard. 2008. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly*, 42(3):375–396.

Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 188–195. Association for Computational Linguistics.

Dana Gablasova, Vaclav Brezina, and Tony McEnery. 2017. Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Language Learning*, 67(S1):155–179.

Dee Gardner and Mark Davies. 2013. A new academic vocabulary list. *Applied Linguistics*, 35(3):305–327.

Beatriz González Fernández and Norbert Schmitt. 2015. How much collocation knowledge do L2 learners have? *ITL-International Journal of Applied Linguistics*, 166(1):94–126.

Stefan Th Gries. 2013. 50-something years of work on collocations. *International Journal of Corpus Linguistics*, 18(1):137–166.

Thomas Herbst, Hans-Jörg Schmid, and Susen Faulhaber. 2014. *Constructions Collocations Patterns*, volume 282. Walter de Gruyter GmbH & Co KG.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36.

Adam Kilgarriff and David Tugwell. 2002. Sketching words. *Lexicography and natural language processing: a festschrift in honour of BTS Atkins*, pages 125–137.

John Lee and Stephanie Seneff. 2007. Automatic generation of cloze items for prepositions. In *Eighth Annual Conference of the International Speech Communication Association*.

Hans Martin Lehmann and Gerold Schneider. 2009. Parser-based analysis of syntax-lexis interactions. *Language and computers studies in practical linguistics*, 68:477.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Yi-Chien Lin, Li-Chun Sung, and Meng Chang Chen. 2007. An automatic multiple-choice question generation scheme for english adjective understanding. In *Workshop on Modeling, Management and Generation of Problems/Questions in eLearning, the 15th International Conference on Computers in Education (ICCE 2007)*, pages 137–142.

Chao-Lin Liu, Chun-Hung Wang, Zhao-Ming Gao, and Shang-Ming Huang. 2005. Applications of lexical information for algorithmically composing multiple-choice cloze items. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 1–8. Association for Computational Linguistics.

Paul Meara. 2009. Simulating word associations in an l2: approaches to lexical organisation. *International Journal of English Studies*, 7(2):1–20.

Haiyan Men. 2017. *Vocabulary Increase and Collocation Learning: A Corpus-Based Cross-sectional Study of Chinese Learners of English*. Springer.

Hilary Nesi. 2011. BAWE: An introduction to a new resource. *New trends in corpora and language learning*, pages 213–228.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language resources and evaluation*, 44(1-2):137–158.

Pavel Rychlỳ. 2008. A lexicographer-friendly association score. *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, 2008:6–9.

Norbert Schmitt. 2010. *Researching vocabulary: A vocabulary research manual*. Springer.

Violeta Seretan. 2008. *Collocation extraction based on syntactic parsing*. Ph.D. thesis, Ph. D. thesis, University of Geneva.

Rita Simpson-Vlach and Nick C Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied linguistics*, 31(4):487–512.

John Sinclair. 1991. *Corpus, concordance, collocation*. Oxford University Press.

Frank A Smadja and Kathleen R McKeown. 1990. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, pages 252–259. Association for Computational Linguistics.

Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *Journal of Machine Learning Research*, 13(Jun):2063–2067.

Alison Wray. 2012. What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics*, 32:231–254.

Jian-Cheng Wu, Yu-Chia Chang, Teruko Mitamura, and Jason S Chang. 2010a. Automatic collocation suggestion in academic writing. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 115–119. Association for Computational Linguistics.

Shaoqun Wu. 2010. *Supporting collocation learning*. Ph.D. thesis, University of Waikato.

Shaoqun Wu, Margaret Franken, and Ian H Witten. 2010b. Supporting collocation learning with a digital library. *Computer assisted language learning*, 23(1):87–110.

A Formulas

The formulas for all the Association Measures (AMs) tried out for Elia can be found at p. 169-170 of (Brezina et al., 2015).

B Supplemental Material

The authors have also released the code, reference data, evaluation results and plots with a Readme document on Github at https://github.com/vishalbhalla/autocoleval to assist the incremental research in the ACL-NLP community. It contains the code for web scraping of both Sketch Engine and FLAX, as well as extracting the filtered collocations for Elia. Since, the Data and Evaluation files for all the three test sets (Sketch Engine, FLAX and Elia) are large, these files can be accessed from https://drive.google.com/open?id=17eydi0KkviG2VxB12l_oNt5LAhuQ6FR0.