Watson Todd, R. (2019) Exploring the direction of collocations in eight languages. *Canadian Journal of Linguistics*

The definitive version of this article was published as Watson Todd, R. (2019). Exploring the direction of collocations in eight languages. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique*, 64(1), March, 146-154. https://doi.org/10.1017/cnj.2018.28

1 Introduction

1.1 Directional Collocation

Within the Firthian tradition in linguistics, and especially since the publication of Sinclair (1991), collocation is viewed as "an integral aspect of linguistic theory" (Barnbrook et al. 2013: 35), yet collocation is largely overlooked in many schools of linguistics. This paper implements a relatively recent measure of directional collocation in corpora of eight different languages to see if there are issues worthy of deeper investigation.

There are two main approaches to investigating collocations. First, the phraseological (Brown 2014) or intensional (Evert 2005) approach treats collocations as falling in the middle of a continuum from idioms to free combinations. Within this approach, collocations may be required to have a non-literal meaning, word spans to identify collocations can be up to four words left and right of the node word, and the identification of collocations is often restricted to combinations of nouns, verbs, adjectives and adverbs. Second, the frequency-based (Brown 2014) or distributional (Evert 2005) approach views collocations as relatively frequent co-occurrences of two words. No constraints on meaning or word types are made in identifying collocations, and the word span is usually one word left or right of the node word. In this paper, I am taking the frequency-based approach associated with corpus linguistics and the key to identifying a collocation is "the extent to which the items appear together more often than we would expect given their individual frequencies" (Brown 2014: 125).

Nearly all previous work on collocation has involved identifying pairs of co-occurring words without considering "whether word1 is more predictive of word2 or the other way round" (Gries 2013: 141). In his original work on collocation in English, Sinclair (1991) distinguished between upward and downward collocations, with upward collocation being where the collocate is a more frequent word than the node, and downward collocation being where the collocate is less frequent. This distinction is important since upward collocation usually highlights grammatical frames, whereas downward collocation highlights semantic issues. An alternative terminology was suggested by Kjellmer (1991) who introduced right-predictive collocations, such as *Pyrrhic victory* where the first word predicts the second but not the other way round, and left-predictive collocations such as *deadly nightshade* where the second word predicts the first (see Michelbacher et al. 2011).

1.2 Measures of Directional Collocation

The standard measures of collocation, such as Mutual Information and z-scores, make no distinction between word1 and word2 and so treat collocations as symmetrical. Thus the asymmetric nature of many collocations has largely been ignored (the only major

exception being the work of Michelbacher et al. 2007; 2011 on directional associations). A few directional measures of collocation were suggested but these were all problematic in some way, and it is only with Gries' (2013) introduction of the ΔP measure that a usable and valid measure has become available.

Gries defines ΔP as

(1) $\Delta P = p(\text{outcome} \mid \text{cue} = \text{present}) - p(\text{outcome} \mid \text{cue} = \text{absent})$

In other words, ΔP is the probability of a word being present given the presence of another word minus the probability of the same word being present without the other word. This allows us to distinguish between right-predictive and left-predictive collocations. A right-predictive collocation will be indicated by a high value for:

(2) $\Delta P_{2|1} = p(\text{word}_2 \mid \text{word}_1 = \text{present}) - p(\text{word}_2 \mid \text{word}_1 = \text{absent})$

A left-predictive collocation will be indicated by a high value for:

(3) $\Delta P_{1|2} = p(\text{word}_1 \mid \text{word}_2 = \text{present}) - p(\text{word}_1 \mid \text{word}_2 = \text{absent})$

An example will show how this works. The words *of* and *course* collocate in English in phrases like *of course* and *in the course of* with an expectation that the left-predictive collocation (*of course*) will dominate. The frequencies of *of* and *course* in the English corpus used in this study are given in table 1. For $\Delta P_{2|1}$, the first probability is the frequency of both words being present divided by the total frequency of *course*; the second probability is the frequency of *of* without *course* divided by the total frequency where *course* is absent (i.e. (273/1140) - (55197/1872038) = 0.210). For $\Delta P_{1|2}$, the first probability is the frequency of both words being present divided by the total frequency of *of*; the second probability is the frequency of *course* without *of* divided by the total frequency where *of* is absent (i.e. (273/55470) - (867/1817708) = 0.004). From this we can see that the left-predictive *of course* is a much stronger collocation than the right-predictive *course of*.

ΔP values range from -1 (where the presence of the cue reduces the likelihood of the outcome, e.g. *they has*) to +1 (where the presence of the cue makes the outcome more likely). In most analyses, collocation measures are applied to those collocations that exist in the corpus being analyzed. Non-occurring pairs are not normally considered. For this reason, negative ΔP values will be much rarer than positive ΔP values. A clear direction of collocation can be found by calculating $\Delta P_{2|1} - \Delta P_{1|2}$ (if positive, this shows a right-predictive collocation; if negative, a left-predictive collocation). Desagulier (2015) provides a clear, detailed explanation of interpreting ΔP values.

2 Focus of Research

As a relatively recent measure, ΔP is yet to be widely used and nearly all applications concern English. It is not clear if the directions of collocations of English are typical of most languages or if different languages have different directional collocation patterns, for instance, in one language most strong collocations might be right-predictive, whereas in another language they might be left-predictive. The purpose of this paper is to conduct

a preliminary analysis of directional collocations in several languages to see if this produces any findings that warrant more detailed investigation.

3 The Analysis

To identify patterns of directional collocations, corpora of several languages are needed. Corpora built on the same principles for numerous languages can be found at the Leipzig Corpora Collection (http://corpora2.informatik.uni-leipzig.de/download.html, see Goldhahn et al. 2012). The following criteria guided corpus selection: the language must be a left-to-right alphabetic language with words separated by spaces, a range of languages falling into different language families should be chosen, and corpus size should be at least 1 million words. From these criteria, the corpora consisting of 100,000 sentences taken from the Internet for eight languages were used. The languages are English, German, Italian and Russian (all Indo-European), Finnish (Uralic), Maltese (Afroasiatic), Indonesian (Austronesian) and Basque (a language isolate). Some potentially relevant typological features of these languages are given in table 2 (based on the World Atlas of Language Structures at http://wals.info/, see Dryer and Haspelmath 2013). Although not ideal corpora given that they are constructed solely from Internet data, these should allow us to conduct a preliminary analysis.

An online program for calculating ΔP values from a corpus was created using word forms as the input (http://jira.org/dp/), and the ΔP values for all immediate collocations with a minimum frequency of 10 were calculated. Various analyses (detailed below) were then conducted to see if any languages have a preference for either right- or left-predictive collocations. The results were subjected to statistical testing using chi-square and Mann-Whitney *U* as appropriate to see if the differences between right-predictive and left-predictive collocations were significant in a given language. Given the number of comparisons made, a level of significance of $p < 0.001$ was used to avoid Type I errors.

4 The Results

The first result concerns the numbers of immediate collocations with a minimum frequency of 10 in each language and this is shown in table 3. The eight corpora are of similar sizes, but there is some variation in the number of collocations identified. This appears to reflect the extent to which a language is synthetic, since more synthetic languages have a greater variety of word forms and so there are fewer common collocations (see Stengers et al. 2011).

Focusing on the 1,000 collocations with the highest ΔP values, we can see if they tend to be more right- or more left-predictive, and the counts for these are shown in table 4. Interestingly, all languages have more left-predictive strong collocations (although for English, for example, the difference is negligible), with five of the languages showing a clear preference for left-predictive collocations.

Separating the left-predictive (i.e. $\Delta P_{1|2}$) and the right-predictive (i.e. $\Delta P_{2|1}$) collocations, we can calculate the average ΔP values for the top 100 and top 500 strongest collocations, and these are shown in table 5. Treating the probabilities as rates of occurrence, to find the average probability we need to use the harmonic mean (the number of items divided by the sum of the reciprocals). Again, most languages show a

clear preference for left-predictive collocations, Indonesian is neutral, and English shows a preference for right-predictive collocations.

Finally, we can focus on those collocations which are unidirectional. For left-predictive collocations, this involves calculating $\Delta P_{1|2}$ - $\Delta P_{2|1}$ (and vice versa for right-predictive). We can then count the number of collocations with ΔP value differences above certain thresholds and the findings for this are shown in table 6. As with the previous analyses, German, Italian and Maltese show a preference for left-predictive collocations. English, on the other hand, has a preference for right-predictive collocations.

To illustrate what these numbers involve, the top 20 unidirectional collocations in English are listed in table 7. It is noticeable that these include only two proper nouns (proper nouns are highly likely to be involved in unidirectional collocations) and that 15 of the collocations include a preposition (a similar pattern is also found for German).

5 Discussion

This is a preliminary speculative study aiming to see whether applying a largely unused measure can lead to insights worthy of detailed investigation. From examining eight languages, it does appear that different languages manifest directional collocation in different ways. Focusing on those points where statistical tests were used, we can summarize the dominant directions of collocations in the eight languages as shown in figure 1, which shows that most languages have a clear preference for left-predictive collocates.

Comparing the directional preferences with the typological features of the eight languages in table 2, the only feature which suggests a relationship with direction of collocations is whether the language is analytic (neutral or right-predictive) or synthetic (left-predictive). It is unclear why this relationship might exist.

For the other typological features, no close relationship with preferred direction of collocation is apparent. This is perhaps highlighted most clearly by adposition types in English and German. The majority of the top 100 directional collocations in both languages include adpositions and both languages use prepositions with noun phrases, yet in the top 100 directional collocations which include prepositions, 53 of 63 are left-predictive in German and 65 of 76 are right-predictive in English reflecting the overall directional preference of each language. In other languages, adpositions are far less common in the top 100 directional collocations. For example, in Italian only 28 include prepositions. Whether other paired sequences of parts of speech are prevalent in the strongly directional collocations in other languages is unclear, and directional collocation analysis using tagged corpora may help.

One potential problem emerging from the findings is that English has a preference contrasting with the majority of the languages. As mentioned earlier, nearly all previous work on directional collocations has focused on English. This emphasis on English is symptomatic of research in several areas of linguistics; a quick Google Scholar search finds that English is the most researched language in reading research, natural language processing, pragmatics and lexis. If English is an outlier among languages (as might be the case for direction of collocations), then the emphasis on English as the focus of research is concerning.

The findings show that different languages do have different preferences for direction of collocations, and that these preferences are realized differently in the various languages raising many questions. Why are most languages left-predictive? Why are prepositions so common in strongly directional collocations in German and English but not in other languages? Why does Indonesian have no clear preference? Is English an outlier language? This paper makes no attempt to answer such questions; rather, it aims to use a ΔP analysis to highlight issues that may be worthy of further consideration.

References

Barnbrook, Geoff, Oliver Mason and Ramesh Krishnamurthy. 2013. *Collocation: Applications and implications*. Basingstoke, Hampshire: Palgrave Macmillan.

Brown, Dale. 2014. Knowledge of collocations. In *Dimensions of vocabulary knowledge*, ed. James Milton and Tess Fitzpatrick, 123-139. Basingstoke, Hampshire: Palgrave Macmillan.

Desagulier, Guillaume. 2015. A lesson from associative learning: Asymmetry and productivity in multiple-slot constructions. <halshs-01184230> Available at: https://halshs.archives-ouvertes.fr/halshs-01184230

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. *The world atlas of language structures online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://wals.info

Evert, Stefan. 2005. The statistics of word cooccurrences: Word pairs and collocations. Doctoral dissertation, Universität Stuttgart.

Goldhahn, Dirk, Eckart, Thomas and Quasthoff, Uwe. 2012 Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC'12), Istanbul, Turkey.

Gries, Stefan Th. 2013. 50-something years of work on collocations. *International Journal of Corpus Linguistics* 18: 137-165.

Kjellmer, Góran. 1991. A mint of phrases. In *English corpus linguistics: Studies in honor of Jan Svartvik*, ed. Karin Aijmer and Bengt Alternberg, 111-127. London: Longman.

Michelbacher, Lukas, Stefan Evert and Hinrich Schütze. 2007. Asymmetric association measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, Borovets, Bulgaria.

Michelbacher, Lukas, Stefan Evert and Hinrich Schütze. 2011. Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory* 7: 245-276.

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

Stengers, Helene, Frank Boers, Alex Housen and June Eyckmans. 2011. Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *International Review of Applied Linguistics* 49: 321-343.

	course present	*course* absent	Total
of present	273	55197	55470
of absent	867	1816841	1817708
Total	1140	1872038	1873178

Table 1 Frequencies of *of* and *course* for calculating ΔP

Language	*Language typology*	*Word order*	*Adposition type with noun phrase*	*Adjective Noun word order*
Basque	Very synthetic	SOV	Postposition	NA
English	Analytic	SVO	Preposition	AN
Finnish	Very synthetic	SVO	Postposition	AN
German	Synthetic	multiple	Preposition	AN
Indonesian	Analytic	SVO	Preposition	NA
Italian	Synthetic	SVO	Preposition	NA
Maltese	Synthetic	SOV	Preposition	NA
Russian	Synthetic	SVO	Preposition	AN

Table 2 Typological features of the eight languages

Language	*Corpus size (no. of words)*	*No. of collocations with a minimum frequency of 10*
Basque	1.9 million	7,946
English	1.9 million	18,776
Finnish	1.4 million	3,597
German	1.8 million	11,590
Indonesian	1.7 million	13,816
Italian	1.8 million	19,082
Maltese	1.4 million	13,117
Russian	1.2 million	5,010

Table 3 Numbers of frequent collocations in eight languages

Language	*No. of right-predictive collocations in the top 1,000*	*No. of left-predictive collocations in the top 1,000*	*Difference between right-predictive and left-predictive collocations*
Basque	465	535	
English	496	504	
Finnish	370	630	$p < 0.001$
German	286	714	$p < 0.001$
Indonesian	476	524	
Italian	336	664	$p < 0.001$
Maltese	415	585	$p < 0.001$
Russian	317	689	$p < 0.001$

Table 4 Numbers of right- and left-predictive collocations in the top 1,000

Language	*Right-predictive collocations*		*Left-predictive collocations*		*Difference between right-predictive and left-predictive collocations for top 100*	*Difference between right-predictive and left-predictive collocations for top 500*
	Harmonic mean top 100	Harmonic mean top 500	Harmonic mean top 100	Harmonic mean top 500		
Basque	0.901	0.500	0.944	0.603	$p < 0.001$	$p < 0.001$
English	0.838	0.537	0.778	0.508	$p < 0.001$	$p < 0.001$
Finnish	0.643	0.266	0.749	0.456	$p < 0.001$	$p < 0.001$
German	0.665	0.342	0.849	0.525	$p < 0.001$	$p < 0.001$
Indonesian	0.839	0.453	0.830	0.454		
Italian	0.813	0.499	0.935	0.645	$p < 0.001$	$p < 0.001$
Maltese	0.939	0.688	0.993	0.819	$p < 0.001$	$p < 0.001$
Russian	0.767	0.279	0.813	0.536		$p < 0.001$

Table 5 Harmonic means of strongest right- and left-predictive collocations

Language	*Right-predictive collocations*			*Left-predictive collocations*			*Difference between right-predictive and left-predictive collocations for 0.75*
	N $\Delta P_{2\|1}$ - $\Delta P_{1\|2}$ > 0.95	N $\Delta P_{2\|1}$ - $\Delta P_{1\|2}$ > 0.90	N $\Delta P_{2\|1}$ - $\Delta P_{1\|2}$ > 0.75	N $\Delta P_{1\|2}$ - $\Delta P_{2\|1}$ > 0.95	N $\Delta P_{1\|2}$ - $\Delta P_{2\|1}$ > 0.90	N $\Delta P_{1\|2}$ - $\Delta P_{2\|1}$ > 0.75	
Basque	12	36	71	21	37	109	
English	10	22	91	6	12	46	$p < 0.001$
Finnish	1	4	18	3	7	29	
German	2	6	22	21	33	72	$p < 0.001$
Indonesian	10	20	62	8	14	47	
Italian	3	8	58	31	57	125	$p < 0.001$
Maltese	14	32	138	89	166	284	$p < 0.001$
Russian	10	16	49	6	16	74	

Table 6 Numbers of unidirectional collocations

Rank	*Collocation*	*Direction*	$\Delta P_{1\|2}$ - $\Delta P_{2\|1}$
1	I reckon	left	0.994
2	year olds	left	0.990
3	accordance with	right	-0.989
4	conjunction with	right	-0.988
5	specialises in	right	-0.981
6	in accordance	left	0.980
7	irrespective of	right	-0.970
8	dispose of	right	-0.970
9	reminiscent of	right	-0.970
10	outskirts of	right	-0.970
11	of Wight	left	0.970
12	per annum	left	0.968
13	in conjunction	left	0.964
14	cater for	right	-0.957
15	specialising in	right	-0.955
16	according to	right	-0.950
17	unable to	right	-0.946
18	New Zealand	left	0.944
19	the foreground	left	0.938
20	the complainant	left	0.938

Table 7 Top 20 unidirectional collocations in English

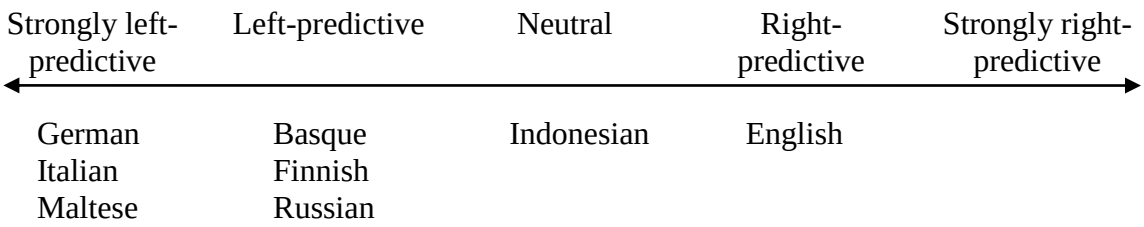

Figure 1 Summary of collocational direction preference in the eight languages