MI-score-based collocations in language learning research: A critical evaluation

Dana Gablasova, Vaclav Brezina and Tony McEnery (Lancaster University, UK)

Introduction and motivation

Formulaic language has occupied a prominent role in the study of language learning and use for several decades (Wray, 2013). Recently an even more notable increase in interest in the topic has led to an 'explosion of activity' in the field (Wray, 2012, p.23). Language learning research (LLR) in both first and second language acquisition has focused on examining the links between formulaic units and fundamental cognitive processes in language learning and use, such as representation of and access to these units in mental lexicon (Wray 2002, 2012, 2013; Ellis et al, 2015). Collocation, a specific unit of formulaic language, holds a prominent position in LLR, having been used in a number of studies on formulaicity in L2 (Schmitt, 2012). The statistical measures for identifying collocations, association measures (AMs), in these studies are of paramount importance as they directly and significantly affect the findings of these studies and consequently the insights into language learning that they provide. One of the most prominent and frequently selected association measure in these studies is the Mutual Information score (MI-score), often referred to as a measure of collocational 'strength' (c.f. Hunston, 2002).

While MI-score has been a useful measure in LLR, there are also several issues related to its use (Gablasova et al. forthcoming 2017). First, the rationale behind the selection of MI-score in studies on formulaic development is not always fully transparent and systematic (González Fernández & Schmitt, 2015) and is often motivated by tradition rather than by specific aims of a given LLR study. Second, alternative measures are rarely considered and their relevance to LLR is not further examined (Gilquin & Gries, 2009). Finally, the application and interpretation of MI-score in LLR suggests that a fuller understanding of the mathematical and linguistic principles on which the measure is based is needed in LLR studies (e.g. an understanding of what type of collocations receive higher MI values and the reasons for this). This understanding would enable a better interpretation of collocational patterns found in L2 production.

Research aims & methodology

This study offers an empirical validation of MI-score based collocations for LLR research. In order to address the above issues, the paper seeks to achieve the following three objectives: i) to place MI-score in the context of other similar association measures and discuss the similarities and differences directly relevant to LLR; ii) to examine the effect of a specific corpus (and register/genre) on the MI values and discuss the implications of the differences in collocational strength measured by MI-score in these corpora; iii) propose general principles for selection of association measures in LLR. The study examines these questions using data from several corpora and sub-corpora (the BNC and its written and spoken sub-corpora, CANCODE and the spoken component of BNC2014 – the newly developed corpus of British English) (see Table 1 for an overview). The whole BNC was included as it has traditionally been used in language learning collocational research as a reference corpus (e.g. Durrant & Schmitt, 2009; Granger & Bestgen, 2014). Five BNC subcorpora were used to investigate the variation in collocational strength inside the BNC. CANCODE and BNC_SP were selected to strengthen the spoken component of the study by looking at more recent corpora of informal speech which are directly comparable with the informal subcorpus (i.e. BNC-Demographic) from the BNC.

Table 1 Overview of (sub) corpora used

Corpus	Size	Representativeness
British National Corpus (BNC)	98,560,118	Written and spoken (10M), diff. registers
BNC_A	15,778,043	Written, academic writing
BNC_N	9,412,245	Written, news
BNC_F	16,143,913	Written, fiction
BNC – Context governed (BNC_CG)	6,196,134	Spoken, formal
BNC – Demographic (BNC_D)	4,234,093	Spoken, informal
BNC – 2014 Spoken (BNC_SP)	4,789,185	Spoken, informal
CANCODE (CANC)	5,076,313	Spoken, informal

We selected three types of collocations representing a range of constructions that commonly appear in language learning collocational research (e.g. Siyanova & Schmitt, 2008; Durrant & Schmitt, 2009; Paquot & Granger, 2012; Granger & Bestgen, 2014, Ebeling & Hasselgård, 2015): verb + complementation (*make* + *sure/decision/point*), adjective + noun (*human* + *beings/rights/nature*) and adverb + adjective (*vitally/very/really* + *important*). Using the selected (sub)corpora, we examined the strength of these collocations using MI-score and contrasted it with two other association measures (Log Dice and t-score) – an example of the results can be seen in Table 2 which illustrates the difference between three collocational measures across different sub(corpora). Special attention was paid to how collocational patterns (i.e. collocational strength) change according to a different genre/register or mode of communication (i.e. written vs spoken language).

Results

The following table provides a brief outline of the results obtained with one of the variables (*make* + complementation) that we investigated as an example.

Table 2 *Make* (lemma), [L0 R2]

	BNC	BNC_A	BNC_N	BNC_F	BNC_CG	BNC_D	BNC_SP	CANC
				MI-score				
sure	6.80	7.09	7.26	5.78	6.90	6.64	6.26	6.92
decision	4.55	3.67	4.07	5.86	6.12	7.91	7.57	8.07
point	3.44	2.92	3.84	3.68	4.11	3.12	3.01	3.93

				t-score				
sure	74.52	13.58	21.45	32.24	28.59	16.92	18.36	22.31
decision	27.59	9.44	9.22	10.45	11.58	5.08	6.82	10.82
point	27.41	9.50	7.78	35.61	13.78	3.43	4.20	6.47
				Log Dice				
sure	9.63	7.60	9.52	9.61	10.68	10.17	10.07	10.61
decision	6.91	6.61	7.16	6.62	8.31	7.05	7.60	8.87
point	6.90	6.65	6.74	6.55	8.52	6.03	6.30	7.30

As can be seen from the table, there are differences between the strength of association between two words according to the measure used and according to the (sub)corpus in which the association was measured. While the difference between measures such as t-score and MI-score was expected, the difference between MI-score and Log Dice deserves further attention as both measures reward similar linguistic properties of collocations (e.g. exclusivity of association). The presentation will discuss the individual results including multiple variables in detail; due to space constraints these could not be fully included in the abstract.

Discussion

With respect to the three research aims, the results revealed:
i) A difference between the three AMs (MI-score, Log Dice and t-score) in identifying the strength of the relationship between words; in particular, the difference between MI-score and LogDice is interesting and has implications for the selection and interpretation of AMs in language learning research.
ii) The variation in the collocational strength between the BNC and its various sub-corpora suggests that large aggregate data such as the whole BNC may hide different distributions of formulaicity across registers and genres as well as across the written/spoken divide.
iii) Following these findings, we propose general principles for the selection of AMs for language learning research. These include the need to understand 1) the mathematical reasoning behind the measure, 2) the scale on which it operates and 3) its practical effect (what combinations of words get highlighted and what gets hidden/downgraded).

References

Ellis, N.C., Simpson-Vlach, R., Römer, U., Brook O'Donnell, M. & Wulff, S. (2015). Learner corpora and formulaic language in second language acquisition. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp 357-378). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139649414.016

Gablasova, D., Brezina, V. & McEnery, T. (forthcoming, 2017). Collocations in corpus-based language learning research: identifying, comparing and interpreting the evidence. *Currents in Language Learning.*

Gablasova, D., Brezina, V., McEnery, T. & Boyd, E. (2015). Epistemic stance in spoken L2 English: The effect of task type and speaker style*. Applied Linguistics* (Advance Access). doi:10.1093/applin/amv055

Gilquin, G., & Gries, S. Th. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*(1), 1-26. doi:10.1515/CLLT.2009.001

González Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have?: The effects of frequency and amount of exposure. *International Journal of Applied Linguistics*, *166*(1), 94-126. doi:10.1075/itl.166.1.03fer

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139524773

Schmitt, N. (2012). Formulaic Language and Collocation. In Chapelle, C. (Ed.), *The Encyclopedia of Applied Linguistics*. New York: Blackwell. doi:10.1002/9781405198431.wbeal0433

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (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. doi:10.1017/S026719051200013X

Wray, A. (2013). Formulaic language. *Language Teaching*, *46*(3), 316-334. doi:10.1017/S0261444813000013