英語コーパス学会 語彙研究会（Japan Association for English Corpus Studies, Lexicology SIG）
Journal of Corpus-based Lexicology Studies, Vol. 6 (2023)
ISSN: 2434-169X

Conflating Directional Association Measures
—A Case Study on NP Constructions—

KAMBARA, Kazuho (Ritsumeikan University)
CHIKA, Taishi (Kansai University)
TAKAHASHI, Norihisa (Ritsumeikan University)
kazy0324@pep-rg.jp
nanou7614@gmail.com
ntakaha@fc.ritsumei.ac.jp

方向をもつ共起強度の統合の試み
—NP 構文を例に—

神原　一帆（立命館大学）
近　大志（関西大学）
高橋　典寿（立命館大学）

Abstract

This paper proposes a method to conflate directional association measures. In measuring associations of collocates, various measures are proposed (e.g., Dice, Jaccard, Mutual Information (MI)). These conventional association measures are said only to reflect limited aspects of collocates. One of the most frequently discussed aspects in collocational analysis is the direction of associations. For instance, when analysts wish to state that the possessive article “my” is likely to occur with the noun “sister”, they need to recognise the two questions regarding the two different directions of association: (i) ‘When the word “my” is used, how likely does the word “sister” collocate?’, and (ii) ‘When the word “sister” is used, how likely does the word “my” collocate?’. Collostructional analysis implements these directions as two types of ΔPs, treated as independent aspects of the association (cf. Gries 2019, 2023). This paper proposes a method to integrate these dimensions. Our approach can account for (i) the sparse distributions of two ΔPs and (ii) the positive correlations between constructions and words more smoothly.

Keywords

Collostructional Analysis, Tupleisation, Conflation, Directional Association

1. Introduction

Various association methods (e.g., Dice, Jaccard, MI (Mutual Information)) are proposed to measure association strengths between words and constructions. These measures are developed because the raw frequency of collocates alone cannot account for the different distributions of each expression. For instance, when analysts observe that the possessive article “my” and the noun “dog” are more likely to cooccur than other nouns in a given corpus, it is hasty to say the article “my” is likely to cooccur with the noun “dog”. This is because analysts must take the distribution of each word into account. When the article “my” cooccurs with the noun “dog”, analysts must compute the conditional probabilities of each combination: (i) the collocating probability of “my” with “dog”, and (ii) the collocating probability of “dog” with “my”. Different association measures are designed to capture different aspects.

In the recent development of collostructional analysis (Gries, 2019, 2023) that investigates the attraction/repelling relations between constructions and words, the tupleisation of different dimensions is proposed. Since many association measures come with advantages and limitations, integrating these values as separate dimensions is proposed (Gries, 2019, pp.394–396). In measuring the associative strength of “my” and “dog”, we can simply plot the noun “dog” in the two-dimensional space of associative probabilities to compare with other nouns. While selecting relevant dimensions depends on each research goal, tupleisation of different associative dimensions can lead to a more accurate understanding of collocations.

However, tupleising multiple dimensions can make the finding quite challenging to generalise, which calls for a method of conflating different association measures. This paper proposes a method to conflate two directions of association between constructions and collocating words by (i) standardising the conditional probabilities of collocates and (ii) measuring distances from the respective reference points. This way, we can obtain a direction-informed rank of desired collocations.

This paper is structured as follows: Section 2 introduces the background and our

research goal: to propose a mathematically sound measure to conflate directional associations. Section 3 describes our proposal in detail using the results of collostructional analysis on a family of noun phrase constructions, henceforth NP constructions (i.e., “the dog”, “my dog”, “Alice’s dog”, “Alice”). Section 4 concludes and discusses possible future developments.

2. Accounting for the various aspects of collocational strengths

This section provides an overview of tupleisation (Gries, 2019, 2023). This method provides complementary information (e.g., type frequency, the bidirectional association between lexical items and constructions, and dispersion) for assessing the validity of association measures computed using collostructional analysis. Section 2.1 identifies the characteristics of collostructional analysis and overviews some recent developments. Section 2.2 reviews our previous work (Kambara & Chika, 2023) to show the merit and the methodological problem within tupleisation.

2.1. Towards a tupleisation of collocational strengths

Several association measures, such as Mutual Information (MI), have been developed to compute the collocational strength between linguistic expressions. Among them, collostructional analysis, initially developed by Gries and Stefanowitsch (2003), would be one of the most effective tools enabling linguists to investigate the mutual association between lexical items and grammatical constructions.

Collostructional analysis is not a single method, but it breaks down into three methods for different purposes (cf. Gries (2019, p.386); Hilpert (2014, p.392)):

i. Collexeme Analysis investigates which lexical items (typically or rarely) occupy a given slot in a single grammatical construction (e.g., keep on V-ing)
ii. Distinctive Collexeme Analysis determines how much words prefer to occur in slots of two (or more) functionally similar constructions (e.g., will V vs. going to V)
iii. Co-varying Collexeme Analysis quantifies how much a pair of lexical items occupy each slot within the same construction (e.g., V1 someone into V2-ing, in

talk someone into buying something)

In conventional analyses, researchers often use a single measure of association by conflating (or confounding) different types of information, such as the type frequency or dispersion of lexical items occurring within/without the construction in question. However, there is a risk that the values of the association vary with sample size and type frequency. For example, type frequency is information that can significantly impact the results of collostructional analysis.

To mitigate this problem, Gries (2019) proposed an approach, tupleisation, that complements the validity of the association measure by considering different information (i.e., tuples) belonging to different dimensions. Gries (2019, p.395) enumerated the following dimensions that should not be ignored:

i. Frequency and effect size in the choice of association measures
ii. The "other" categories in both the rows and the columns of the traditional 2 × 2 tables
iii. The directions of association/repulsion of the two elements involved
iv. Frequencies from whole corpora, regardless of the elements' dispersions

2.2. A way to conflate directional associations

This section reviews the advantages and challenges of tupleisation. Gries (2019) proposed a research programme, tupleisation, to emphasise the importance of accounting for the different aspects of collocational strength. Tupleising association measures can aid the limitations of conventional collocation studies. However, tupleising various association measures can make the findings challenging to generalise, which calls for a well-informed conflation measure. For a case study, we review Kambara and Chika (2023), which was largely inspired by Löbner (2011) and Glass (2022), to show how conflation can be a challenge.

2.2.1. Tupleisation in action

Kambara and Chika (2023) conducted a collostructional analysis on NP construction

to show the effectiveness of tupleisation. Inspired by Löbner (2011), Glass (2022) observed that relational nouns, such as body-part terms (e.g., "arm") and kinship terms (e.g., "sister"), are more likely to be realised as the head of possessive constructions (e.g. [NP of NP], [N's N]) than sortal nouns (e.g., "dog", "apple"). Previous studies assumed that there is a binary distinction between relational nouns and non-relational nouns (cf. de Bruin & Scha, 1988; Barker, 2011), which does not necessarily accord well with actual observations (e.g., "my dog").

Despite Glass's promising outlook, Kambara and Chika (2023) pointed out two limitations: (i) the scope of the analysis and (ii) the need for refining association measures. The first limitation corresponds to the relatively limited types of grammatical constructions. Löbner (2011) originally proposed a distribution-based classification of nouns using relationality (i.e., whether the noun's referent is determined by its own) and uniqueness (i.e., whether the noun's referent is unique). Relationality is measured by the likelihood to occur in possessive constructions (e.g., "my pet"), while uniqueness in bare construction (e.g., "Alice"). To support Löbner's account, the scope of the analysis must extend to the whole NP construction. The second limitation revolves around the issues pointed out in Section 2.1. Glass measured the association between a possessive construction and given words by computing the proportions of collocating nouns, which cannot account for the other associative direction (i.e., the attraction from the word to the construction).

An extensive collostructional analysis (more specifically, collexeme analysis) was conducted to overcome these two challenges. Kambara and Chika semi-manually tagged the data extracted from the BNC component of the Treebank Semantic Parsed Corpus (TSPC) (Butler, 2022) to categorise NP Constructions into four types (i.e., Cx1, Cx2, Cx3, Cx4), as summarised in Table 1.

Table 1. Basic statistics of NP constructions

NP Construction	Token Freq (Relative Freq)	Type Freq (Relative Freq)
Cx1 (e.g., [the X of Y])	844 (0.080)	485 (0.103)
Cx2 (e.g., [my X])	810 (0.077)	532 (0.113)
Cx3 (e.g., [a X], [the X])	3,160 (0.299)	1,367 (0.291)
Cx4 (e.g., [X])	5,746 (0.544)	2,310 (0.492)

Then, using the R script "Coll.analysis 4.0." to perform collostructional analysis (Gries, 2022a), they computed the associative directions: (i) The attraction from the word to the construction, and (ii) from the construction to the word. These attractions are referred to as $\Delta P_{w \to cx}$ and $\Delta P_{cx \to w}$, respectively (where w stands for the word, cx for the construction, and the arrow for the associative direction). The result of the analysis is visualised as a scatterplot, as shown in Figure 1. The obtained result was largely consistent with Glass's findings, and they pointed out that more detailed analysis is needed to neatly classify the sortal nouns (cf. Cx3) and proper names (i.e., Cx4).

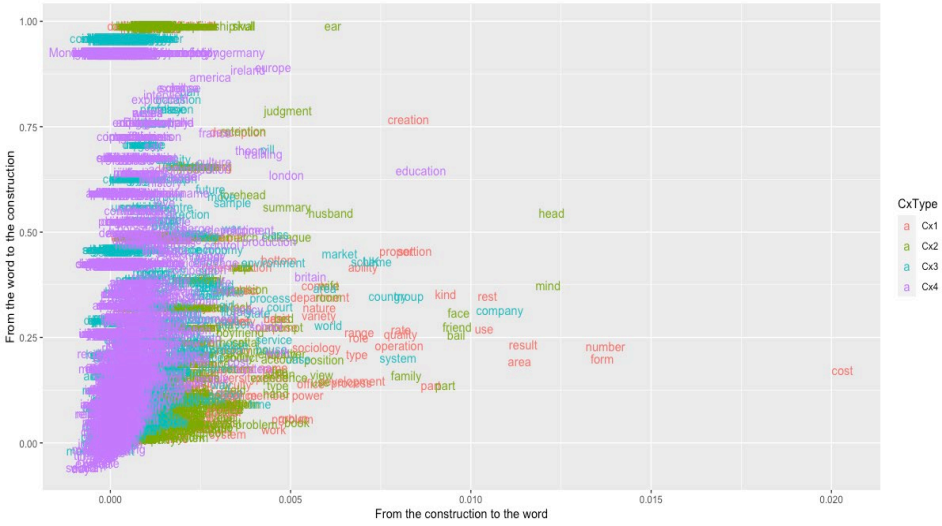

Figure 1. Tupleisation of directional association measures

2.2.2. Issues in conflating dimensions

Tupleising multiple directions of associations allows analysts to consider the collocates carefully. However, how analysts should conflate their desired association measures is controversial, to say the least. We review two approaches to conflating these dimensions by Gries (2019) and Kambara and Chika (2023), along with their limitations.

Kambara and Chika (2023) propose an intuitive approach to conflate dimensions using medians of each dimension. For each dimension, each value was evaluated if it exceeds the median or not to obtain the typical instances of each construction. Then, the typicality of collocates is categorised into three categories: High (i.e., Both attractions exceed the medians), Medium (i.e., One of the attractions exceeded at least one median), and Low (i.e., Otherwise). Then, for each typicality category, all words are sorted by their raw frequency. Though highly intuitive, their approach has two challenges: (i) the use of median is not properly justified, and (ii) it is odd since the sorting process is evaluated irrelevant to the constructed association space.

Gries (2019, pp. 406–409) proposes a more rigorous approach to conflate various dimensions in the following three steps:

i. Choose the dimensions of information to include.
ii. Convert them all to an equal range (e.g., transforming values to fall into the interval from 0 to 1).
iii. Represent the words by points and measure the Euclidian distance from the origin (i.e., 0).

Gries' approach is more sophisticated because it involves the standardisation of values and the use of Euclidian distance from the origin, which is defined by the constructed association space. However, the standardisation process can be tricky since the distribution of a dimension can be sparse. For instance, as shown in Figure 1., while the values of $\Delta P_{w \to cx}$ range from 0 to 1, those of $\Delta P_{cx \to w}$ range from 0 to 0.02, which suggests a directional association measure can be more or less dispersed than the other. Values of $\Delta P_{w \to cx}$ tend to be higher with low frequency items. The dispersions of probability can affect the computation of distances. This approach could distort the ranking of collo-

cating words because of the overdispersed measures. Though Gries (2019) incorporates other dimensions other than $\Delta P_{\{cx \rightarrow w,\ w \rightarrow cx\}}$, a similar challenge can arise in conflating dimensions.

To overcome the limitations of these two approaches, we need an approach to conflate multiple dimensions of associations, using the constructed association space while caring for the typicality of collocates. The following section aims to present a procedure to conflate directional associations. Note that we only deal with directions of association for simplicity, though nothing stops analysts from including other dimensions.

3. Proposals

3.1. A weighed Euclidian distance as an association measure

This section explains an approach to measuring Euclidian distances in the given association space. Our proposal can be summarised in the following four steps: (i) standardise the values of each associative dimension in z-scores, (ii) compute the Euclidian distance from the origin coordinate (i.e., (0, 0)), (iii) compute the interior angle from the origin to the data point, then (iv) compute the product of (ii) and (iii). Our proposal can mediate the sparse distributions of different dimensions. In the following, each step is explained in detail.

To measure the distances from the origin, standardising association measures is needed. As observed in Figure 1, low-frequency items are assigned with high $\Delta P_{w \rightarrow cx}$. If a lexical item only occurs once in the given observation, the association between the investigated construction and the word is assigned 1 (being highly likely to attract the construction). On the other hand, ranges of $\Delta P_{cx \rightarrow w}$ tend to be smaller especially when the construction's frequency is relatively high. For this reason, we standardise each value using z-scores, which can smoothen the distribution of $\Delta P_{\{cx \rightarrow w,\ w \rightarrow cx\}}$ to some extent. The value of z-scores represents the distance between that raw score and the population mean in units of the standard deviation, in which z is negative when the raw score is below the mean and positive when above.

Then, we compute the Euclidian distance from the origin coordinate $(0, 0)$, an average point where the mean of $\Delta P_{cx \rightarrow w}$ and $\Delta P_{w \rightarrow cx}$ meet. Euclidian distance between the origin "o" and the data point "l" (i.e., the coordinate of the plotted lexical item) is computed by

the following equation. This way, we can measure the distance from the origin to the desired data points in the given association space like Figure 1.

$$|o - l| = \sqrt{(o - l)^2 \cdot (l - o)^2}$$

Using Euclidian distances from the origin can distort the attraction/repelling relations between lexical items and constructions. As discussed, the range of $\Delta P_{w \to cx}$ tends to be higher than that of $\Delta P_{cx \to w}$ when the low-frequency items are involved. If a lexical item occurs only once in the given construction and the frequency of the construction is relatively high, Euclidian distances of collocates are highly influenced by the low-frequency items. For instance, the distance from the origin increases when some items have high $\Delta P_{w \to cx}$. As observed in Figure 1, some items are concentrated on the y-axis of the plot and do not spread across the x-axis.

To mediate the different distributions of $\Delta P_{cx \to w}$ and $\Delta P_{w \to cx}$, we multiply the Euclidian distance of the plotted lexical item with the cosine value of interior angle θ. This operation allows collocates with high standardised $\Delta P_{w \to cx}$ and low standardised $\Delta P_{cx \to w}$ to be evaluated lower than those with high standardised $\Delta P_{w \to cx}$ and high standardised $\Delta P_{cx \to w}$. If both values of standardised $\Delta P_{\{cx \to w,\ w \to cx\}}$ include positive values, the value of θ is not larger than $\pi/2$. The evaluation method can be summarised as the following equation (where "x" denotes the final weighed Euclidian distance, "$dist(o, l)$" the raw Euclidian distance from the origin "o" to the data point "l", "θ" the interior angle, and "$cos(\theta)$" the cosine value of "θ").

$$x \coloneqq dist(o, l) \cdot cos(\theta)$$

3.2. A case study: NP constructions

The following subsections report the results. Using the evaluation methods in Section 3.1., we obtained a list of frequent collocates of each NP construction. Since our analysis yields a straightforward result, interpretations improved. Section 3.2.1. describes the data extraction methods and Section 3.2.2. presents the results and discussions.

3.2.1. Data extraction and processing

For a case study, we chose the data provided by Kambara and Chika (2023), which is a result of collexeme analysis on all variants of NP Construction (i.e., Cx1, Cx2, Cx3, and Cx4 in Figure 1). This data was originally taken from Treebank Semantic Parsed Corpus (TSPC) (Butler, 2022), and all instances are categorised into four categories, as already described in Table 1. The total record of raw data is 4,694 (485 tokens for Cx1, 532 tokens for Cx2, 1,367 tokens for Cx3, and 2,310 tokens for Cx4), accompanied by the following information. Except for the type of NP construction, other types of data (i.–x.) are implemented on the Coll.analysis 4.0. (Gries, 2022a).

i. Lemmatised nouns that occurred at least once as the head of an NP construction
ii. Frequency in the NP construction
iii. Frequency in other constructions (i.e., Frequency in the corpus minus the value of ii.)
iv. The relation of attraction (e.g., attraction vs. repelling)
v. The log-likelihood ratio of collocates
vi. Pearson residuals of collocates
vii. The log-odds ratio of collocates
viii. MI of collocates and the construction
ix. The noun's attraction of the construction (i.e., $\Delta P_{w \to cx}$)
x. The construction's attraction of the noun (i.e., $\Delta P_{cx \to w}$)
xi. The type of NP construction (i.e., Cx1, Cx2, Cx3, Cx4)

We devised a function to compute the weighed distances conflating the directions of associations (i.e., $\Delta P_{\{cx \to w,\ w \to cx\}}$), which is made available on Open Science Framework (OSF) along with the processed data. We compared the obtained results with Kambara and Chika (2023) both qualitatively and quantitatively. All computations were conducted using R (R Core Team, 2022), with a family of ggplot2 (Wickham & Grolemund, 2016) for visualisation.

3.2.2. Results & Discussion

The result of the analysis can be summarised in Table 2, which lists the top ten lexical items with high adjusted distances. The analysis yielded partially different results from the one presented in Kambara and Chika (2023), shown in Table 3. As discussed, they failed to mediate the overdispersed distribution of $\Delta P_{cx \to w}$ in tupleising directional association measures. We report and discuss some implications of our proposal by comparing our analysis with the conventional ones.

Table 2. Typical nouns of each NP construction (sorted by the adjusted distance)

	Cx1	Cx2	Cx3	Cx4
1	cost	head	company	education
2	number	mind	group	Britain
3	form	face	system	London
4	result	friend	country	work
5	area	bail	scheme	Europe
6	rest	part	UK	production
7	use	family	market	school
8	kind	ear	world	training
9	part	husband	area	system
10	creation	wife	case	theory

Table 3. Typical nouns obtained by Kambara and Chika (2023)

	Cx1	Cx2	Cx3	Cx4
1	number	head	group	Germany
2	form	mind	country	fiction
3	result	face	scheme	custody
4	area	friend	UK	monopoly
5	rest	bail	market	warming
6	use	ear	area	correspondent
7	kind	husband	class	agriculture
8	creation	wife	environment	Lister
9	proportion	room	process	California
10	set	judgment	pill	Japan

Based on Table 3, Kambara and Chika reported that (i) Cx1 usually associates with fixed expressions (e.g., "the number of …"), (ii) Cx2 with relational nouns (e.g., "my head"), (iii) Cx3 with sortal nouns (e.g., "country"), and Cx4 with abstract and proper

nouns (e.g., “Germany”, “fiction”). Though some nouns on Table 3 are not identical to those on Table 2, the refined sorting with Euclidian distance invites similar generalisations.

In Cx1, the effect of fixed expressions (or multiword expressions) is quite strong, and the popularity of event-denoting nouns (e.g., “operation”, “development”) decreases. In Cx2, most words are shared, suggesting that the powerfulness of possessive constructions to identify relational nouns. In Cx3, though most nouns are changed, the popularity of sortal nouns (e.g., “company”, “country”) is still strong in this variant. However, like conventional sorting, note that some prototypical sortal nouns (e.g., “dog”) do not appear as one of the most frequent collocates. Finally, in Cx4, the popularity of proper nouns decreased, and the popularity of other sortal nouns increased.

Since these qualitative findings are far from conclusive, we also investigated the quantitative correlation between conventional sorting and the present proposal. Figure 2 shows the correlation between the proposed and the conventional ranks in Kambara and Chika (2023). As confirmed from the clouded scatterplot, the proposed rank is quantitatively different from the conventional one proposed by Kambara and Chika (2023). Though our proposal invites further investigations of its effectiveness, the present conflation technique should be treated as a better alternative since it involves a careful evaluation of the typicality of each collocate. The distribution of these rankings seems like overlapping mirrored letter “L”s, which is probably caused by the categorical sorting technique. In this sense, frequency-based sorting is too naïve to capture the complex nature of collostructional analysis.

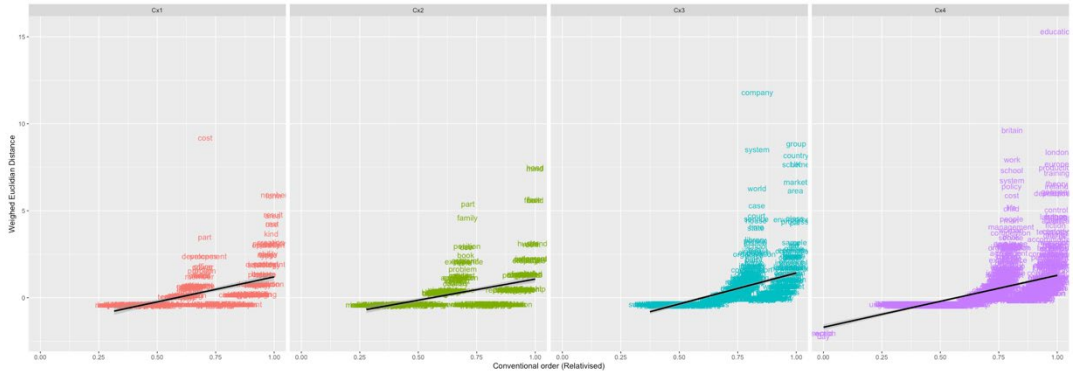

Figure 2. The scatter plot of the conventional ranks and proposed ranks

As opposed to tupleisation, a way of simplification in collostructional analysis is proposed (Gries, 2023, pp.358–365). In Coll.analysis 4.0. (Gries, 2022a), Pearson residuals are recorded to observe the collocational preference of a construction. Simply put, Pearson residuals are the differences between observed and expected values (cf. Gries, 2021, pp.184–185). A positive Pearson residual value corresponds to a positive correlation while a negative one to a negative correlation. Interpretation of Pearson residuals can be carried out in the same way as our proposed weighed Euclidian distance.

We created another scatterplot with Pearson residuals on the x-axis, as shown in Figure 3, which includes fitted regression lines with confidence intervals. Intuitively, the positive correlation between the weighed Euclidian distance and Pearson residuals was confirmed. The construction type with more instances (i.e., Cx3, Cx4) yielded steeper slopes. Our approach can capture some typical collocates of a construction, which allows a more straightforward interpretation of the obtained results.

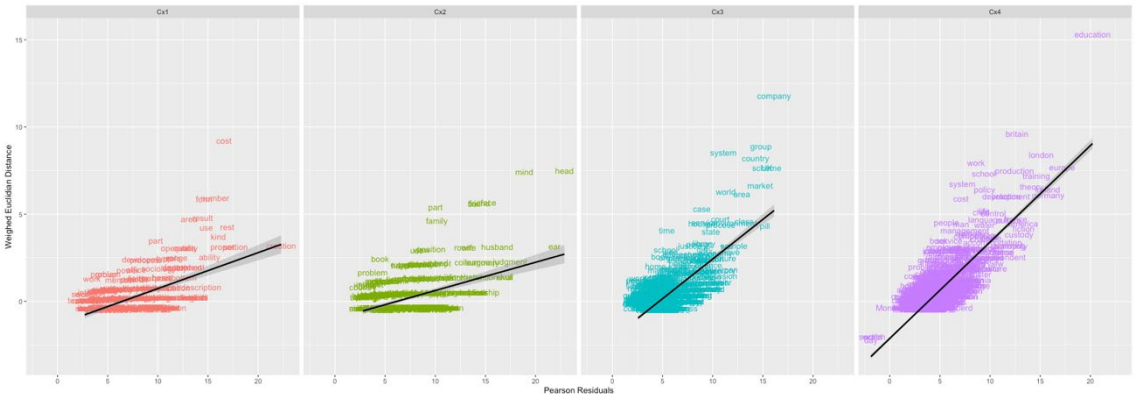

Figure 3 Correlation between Pearson residuals and weighed Euclidian distance

4. Conclusion

This paper sketched a framework to conflate the directional association measures in collostructional analysis. Our proposal consists of four steps: (i) Standardise the values of each associative dimension in z-scores, (ii) compute the Euclidian distance from the origin coordinate (i.e., (0, 0)), (iii) compute the interior angle from the origin to the data point, then (iv) compute the product of (ii) and (iii). These computations should be treated

as a better conflation technique than the conventional one (Kambara & Chika, 2023) because it defines the attraction of collocates based on the association space constructed by the different directions of associations and allows easier interpretations of attraction/ repelling relations.

Two issues remain unsolved. Firstly, the effectiveness of our sorting is still yet to be clarified. Though the preliminary statistical analysis revealed a discrepancy from the conventional analysis, its quantitative and qualitative aspects need to be closely examined. Secondly, tupleisation of associative dimensions does not stop at directional association measures. As Gries (2019, 2023) show, including various dimensions allows more precise generalisations. For this reason, conflation techniques should be generalisable to any number of dimensions in principle. However, our proposal is only applicable to the two-dimensional space of directional association measures.

Data Availability

Analysed data and R scripts are uploaded on Open Science Framework (OSF).

Bibliography

Barker, C. (2011). Possessives and relational nouns. In Maienborn, C., Heusinger, K., & Portner, P. (Eds.), *Semantics: An International Handbook of Natural Language Meaning* (pp. 1109–1130). Mouton de Gruyter.

Butler, A. (2022). The Treebank Semantics Parsed Corpus (TSPC). Hirosaki University. (Accessed 9 June 2023). http://www.compling.jp/ajb129/ts.html

de Bruin, J. & Scha, S. (1988). The interpretation of relational nouns. In *26th Annual Meeting of the Association for Computational Linguistics* (pp. 25–32).

Glass, L. (2022). Quantifying relational nouns in corpora. *English Language & Linguistics*, *26*(4), 833–859.

Gries, S. T. (2019). 15 years of collostructions: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics*, *24*(3), 385–412.

Gries, S. T. (2021). *Statistics for Linguistics with R: A Practical Introduction* (3rd edition). Mouton de Gruyter.

Gries, S. T. (2022a). Coll.analysis 4.0. A script for R to compute perform collostructional analyses.

Gries, S. T. (2022b). Toward more careful corpus statistics: Uncertainty estimates for frequencies, dispersions, association measures, and more. *Research Methods in Applied Linguistics*, *1*(1), 100002. https://doi.org/10.1016/j.rmal.2021.100002.

Gries, S. T. (2022c). What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies*, *5*(1), 1–33.

Gries, S. T. (2023). Overhauling collostructional analysis: Towards more descriptive simplicity and more explanatory adequacy. *Cognitive Semantics*, *9*(3), 351–386.

Gries, S. T. & Stefanowitsch, A. (2004a). Co-varying collexemes in the into-causative. In Achard, M. & Kemmer, S. (Eds.), *Language, Culture, and Mind* (pp. 225–236). CSLI.

Gries, S. T. & Stefanowitsch, A. (2004b). Extending collostructional analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics*, *9*(1), 97–129.

Hilpert, M. (2014). Collostructional analysis: Measuring associations between constructions and lexical elements. In Glynn, D. & Robinson, J. A., (Eds.), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy* (pp. 391–404). John Benjamins.

Kambara, K., & Chika, T. (2023). Toward a corpus-based identification of nominal relationality and uniqueness: A constructionist approach. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation* (pp.661–669).

Löbner, S. (2011). Concept types and determination. *Journal of Semantics*, *28*(3), 279–333.

Stefanowitsch, A. & Gries, S. T. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209–243.

Stefanowitsch, A. & Gries, S. T. (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory*, *1*(1), 1–43.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Wickham, H., & Grolemund, G. (2016). *R for Data Science: Import, Tidy, Transform,*

Visualize, and Model Data. O'Reilly Media, Inc.