Lexical Networks Constructed to Correspond to Students' Short Written Responses: A Quantum Semantic Approach

Ismo T. Koponen[1(✉)], Ilona Södervik[2], and Maija Nousiainen[1]

[1] Department of Physics, University of Helsinki, Helsinki, Finland
ismo.koponen@helsinki.fi
[2] Centre for University Teaching and Learning (HYPE), University of Helsinki, Helsinki, Finland

Abstract. A simple method to construct lexical networks (lexicons) of how students use scientific terms in written texts is introduced. The method is based on a recently introduced quantum semantics generalization of a word-pair co-occurrence. Quantum semantics allows entangled co-occurrence, thus allowing to model the effect of subjective bias on weighting the importance of word co-occurrence. Using such a generalized word-pair co-occurrence counting, we construct students' lexicons of scientific (life-science) terms they use in their written responses to questions concerning food chains in life-science contexts. The method allows us to construct ensembles of lexicons that probabilistically simulate the variability of individual lexicons. The re-analyses of the written reports show that while sets of top-ranking terms contain nearly the same terms irrespective of details of the method used to count co-occurrences, the relative rankings of some key-terms may be different in quantum semantic analysis.

Keywords: Lexical networks · Quantum semantics · Education

1 Introduction

Science learning requires from students adoption of specific type of scientific language, which has its own vocabulary, semantics and syntax; the language of science, as some researchers call it (see e.g., [1]. Learning the language of science, its terms and their correct use is central in the process of learning science and scientific thinking [2,3] as well as in building scientific claims and communicating scientific knowledge [4]. To learn the language of science, students must learn a kind of new lexicon (i.e. lexical network) consisting of specific words, terms

H. Cherifi et al. (Eds.): COMPLEX NETWORKS 2016 2022, SCI 1077, pp. 137–149, 2023.
https://doi.org/10.1007/978-3-031-21127-0_12

and concepts and understand, for example, new phenomena by using that new lexicon. Consequently, to study how language of science is learned, constructing lexicons of students' use of scientific vocabulary and terms in a written text is an important problem in many settings of educational research [5–7].

In this study, we introduce a simple method to investigate how students use scientific terms in written texts, i.e. how they master the basic vocabulary of scientific language. Such a method is needed for three reasons. First, students' written answers are often too short and sparse to allow individual responses being analysed, in which case only aggregated answers yield to analysis. Second, we need an approach that allows us to estimate the extent of individual variations in lexical networks. Third, in educational applications simple enough methods for text analysis are needed because advanced methods of text analysis [8–11] require expertise and are thus not likely to be adopted.

The method introduced here is designed to meet these three demands. The method is based on the recently introduced quantum semantic approach allowing entanglement -kind interdependence of word-pairs as a generalization of word-pair co-occurrence counting, and thus, allows to model a kind of subjective bias in word co-occurrence [12–14]. Such a subjective bias refers to a possibility that different readers or writers of the text may lay a different emphasis on co-occurrence of words in a sentence regarding meaning of words (for a more detailed discussion, see refs. [12–14]). Due to the possibility of taking into account subjective bias we are able to simulate probabilistically the range of individual variations of lexicons corresponding short texts. Consequently, this study contributes to developing a simple method to investigate students' lexicons of language of science for educational applications.

To demonstrate the viability of the method we use as an example first-year life science students' written answers to tasks related to the role of photosynthesizing plants in the ecosystem from the viewpoints of food chains. The same topic was investigated in our previous study [5], where we utilized more conventional network methods [7] to analyze the students' responses but only in the form of an aggregated lexicon. In the previous study [5], the analysis was hampered in many ways by the sparsity of texts given as responses; each text contained only few key-words and sentences were too short to allow detailed analysis to take sentence structure into account and perform co-occurrence counts (in short sentences, co-occurrences are too rare). Therefore, it was possible to analyze only an aggregated lexicon, aggregating about 100 individual texts. Aggregation, however, always creates artefactual connections, and it is difficult to estimate how severely such artefacts bias inferences about the significance of key-terms.

The approach presented here is still based on aggregation of lexicons, but now utilizing quantum semantics [12–14], which now allows us to to estimate variability in linking key-terms in individual lexicons, owing to possibility of entanglement as a model of subjective bias. Consequently, we can then construct ensembles of lexicons that probabilistically simulate the variability of individual lexicons. The re-analysis of the written reports studied previously [5] show that

while sets of top-ranking terms contain nearly the same terms irrespective of details of the method to count co-occurrences, the relative rankings of some key-terms may be different in quantum semantic analysis.

2 Methods and Materials

The data analyzed here comes from our previous study [5], where the written texts analyzed were short answers by first-year university students' of life sciences to questions regarding photosynthesis and its role in food chains in ecosystems, a topic which is of major importance in studying life sciences. Data were collected using open-ended questions that required application of basic conceptual knowledge. The tasks and the instruction were: explain the role of plants in the ecosystem from the food chain's point of view. The tasks entailed an understanding of basic biological phenomena, the most important being photosynthesis. The participants were 150 first-year university-level life science students. The answers to the questions were short and restricted to six lines. Of all answers 100 were long enough for analysis. For this study, no new data were collected; only the data from a previous study [5], made available in fully anonymized, transcribed and lemmatized form is used here.

Other details of the task, data collections, and issues related to it of the educational aspects and uses of the topic are reported elsewhere [5] and are not of further interest here. In what follows, we focus on developing and discussing the methods of analysis that can be used in analyzing sparse texts in educational settings.

2.1 Word Co-occurrence Counts and Concurrence

The first step in transforming the written texts into semantic networks, in the form of lexicons of terms, consisted of splitting the sentences into clauses, in order of their appearance, and after that, recognizing the nouns and within them, term-like words. Note that sentence structure was preserved, instead of using fixed lengths of text sequences as is often done in co-occurrence counting. Term-like words that appeared only once were discarded, and from the remaining set, about 70 were chosen for closer attention. The set of potentially interesting key-words were first selected as basis of their frequency of appearance and second, on basis of their importance for the topic in question. Words and terms deemed to be irrelevant or too commonplace of being further interest we discarded (for details, see [5]). Consequently, the selection process is quite simple and misses many finer points, but contains basic and robust elements commonly identified in analyzing speech and writing.

The second step consisted of finding the co-occurrence statistics for the 70 selected terms. To do so, we utilized quantum semantics approach [12–14], which can be considered a generalization of contingency based analysis of co-occurrence and in particular, as related Yule's Y-factor as a measure of contingency [15,16] (see also Appendix A).

To obtain concurrence Q (degree of entanglement) between words A and B in a text, we count four different frequencies of co-occurrence: the frequency n_{11} that A and B both occur in a given block of clauses at least once, n_{00} that neither A and B occurs, n_{10} that A occurs at least once but B does not occur, and n_{01} that A does not occur but B occurs at least once. With these frequencies, the concurrence Q as a measure of entanglement is given as [12,13].

$$Q = Q_0\sqrt{(1-\Theta R)}\,, \quad 0 \le Q \le 1\,, \tag{1}$$

where $-1 \le \Theta \le 1$ is the (compound) phase factor (see Appendix A for details and derivation), is taken here as a free parameter to account for a kind of subjective bias' [12–14]. In that, the subjective bias refers to various subjective ways to read meanings into the co-occurrence of terms A and B, i.e. to emphasize importance of co-occurrence differently [12–14]. The prefactor $Q_0 = 2\sqrt{n_{01}\,n_{10} + n_{00}\,n_{11}}$ corresponds to concurrence obtained with $\Theta = 0$ and takes into account the marginals of conditional probabilities $n_{01}\,n_{10}$ and $n_{00}\,n_{11}$ (note that rarely occurring pairs are here of no interest). The factor $R = 2\sqrt{n_{11}\,n_{00}\;n_{10}\,n_{01}}/(n_{11}\,n_{00} + n_{10}\,n_{01})$ is the ratio of the geometric mean of frequencies $n_{11}\,n_{00}$ and $n_{10}\,n_{01}$ to their arithmetic mean. The entanglement is now possible only in cases $R \neq 0$.

2.2 Constructing Lexicons

The concurrence $Q(i,j) = Q(j,i)$ of all pairs of terms i and j is obtained by using Eq. (1) for all pairs of terms of interest in students' responses (aggregated). The pairwise values $Q(i,j) = Q(j,i)$ are then used to form the weighted (symmetric) adjacency matrix Q with elements $[\mathrm{Q}]_{ij} = Q(i,j)$ describing the connectivity of terms in the lexicon.

Aggregated lexicon is then used as to generate individual lexicons. The existence and concurrence of links between the terms in an individual lexicon is modelled using parameter Θ as a random variable, to create three cohorts of interest: cohorts corresponding to low $(0.5 \le \Theta \le 1.0)$, high $(-1.0 \le \Theta \le -0.5)$ and averaged values $(-1 \le \Theta \le 1)$, respectively, to create ensembles of possible model lexicons. The model lexicons contain thus the same terms (nodes) as the individual texts but connections are predicted on the basis of aggregated lexicon by using random factors Θ. These ensembles provide then information of bounds of variation allowed in individual lexicons even if co-occurrence frequencies n_{01}, n_{10}, n_{00} and n_{11} are fixed.

2.3 Finding Key-Terms

A correlation matrix describing how nodes are correlated can be introduced by utilizing an exponential adjacency kernel (matrix exponential transformation) [17] to define a correlation matrix $\mathbf{G}$ of the form

$$\mathbf{G} = \mathbf{D}^{-\frac{1}{2}}\ \exp[\,\beta\mathbf{Q}\,]\ \mathbf{D}^{-\frac{1}{2}}\,, \tag{2}$$

where $\mathbf{D} = \mathrm{Diag}[\exp[\beta\mathbf{Q}]]$ is a diagonal matrix. Here, correlations refer to"positions" of nodes in the network, position meaning how through different paths of different lengths nodes can be reached; the more easily the nodes can be reached from each the larger their correlation (for a detailed discussion, see ref. [17]).

The elements of $[\mathbf{G}]_{ij}$ of the correlation matrix can be shown to be directly related to the covariance of values of nodes i and j in the network [17], providing a way to find the key-nodes based on correlations. Towards this end, we define a correlation centrality as an off-diagonal sum

$$\Gamma_i = \sum_{j \neq i} [\mathbf{G}]_{ij} \,, \tag{3}$$

which closely resembles communicability centrality [18,19]. The correlation centrality is used here as a basis to define key-terms and their rankings.

2.4 Similarity Comparison

The similarity of lexicons can also be operationalized based on correlation centrality Γ_i. Forming for lexicons L and L' centrality vectors $\bar{\Gamma}$ and $\bar{\Gamma}'$ consisting of centralities Γ_i and Γ_i' of nodes, respectively, we can define the similarity S_0 as a dot-product [20,21]

$$S_0[\mathrm{L}||\mathrm{L}'] = \bar{\Gamma} \cdot \bar{\Gamma}' \,, \tag{4}$$

which projects the centrality vectors to each other. For comparison we use normalized similarity (called cosine-similarity) given by $S_N[\mathrm{L}||\mathrm{L}'] = S_0[\mathrm{L}||\mathrm{L}']/(\,|\bar{\Gamma}|\,|\bar{\Gamma}'|\,)$, with range limited to $0 \leq S_N \leq 1$. The similarity S_0 takes into account the size of the lexicons, while S_N does not depend explicitly on size; high similarity lexicons have high values of both similarity measures.

3 Results

The frequency of occurrences of words and terms in a text is the simplest thinkable measure characterizing the lexicons. An obvious step further is to perform co-occurrence counts of word pairs, to be used as a basis to construct lexicons, where co-occurrence correlations are represented as links between pair of words. Such a lexical network of connections constructed as an aggregate of connection in all sentences in 100 written sparse texts is shown in Fig. 1. The size of the nodes is proportional to correlation centrality as defined by Eq. (3). The network shown is for a moderate entanglement with $\Theta = -0.5$. Symbols denoting terms are explained in Table 1.

In the aggregated lexicon shown in Fig. 1, the six most important terms are: "plant" (P), "food chain" (Fc), "energy" (E), "producer" (p), "organism" (O) and "nourishment" (N) (for other terms, see Table 1). Different choices for factor

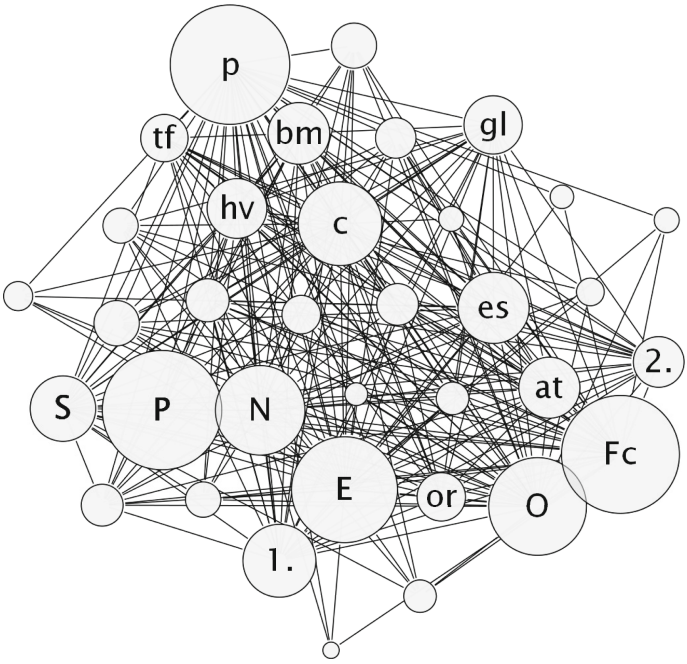

Fig. 1. The aggregated lexicon (lexical network) of key-terms in students' responses to questions about the role of plants in the ecosystem from the viewpoint of the food chain. The symbols denoting key-terms are explained in Table 1. The links that exceed a threshold-value of 0.30 are shown. The size of the nodes correspond to their correlation centrality as defined in Eq. (3).

Θ produce different entanglements, and thus different rankings, but five of the six top terms are always present in the top-five cohort.

The texts analyzed here are sparse, with quite short sentences (mostly statements), so that formation of lexicons corresponding to an individual text is in most cases impossible. However, the aggregated lexicons shown in Fig. 1 now provide information on how probable a given connection is in the whole set of individual texts. Note, however, that only a fraction of the terms might be present in individual texts. To predict the existence of connections between the terms in an individual lexicon, we use random phases, distributed into three cohorts corresponding to low values of concurrence (Low=L, with $0.5 \leq \Theta \leq 1.0$, high values of concurrence (High=H, with $-1.0 \leq \Theta \leq -0.5$, and total range (Total=T, with $-1 \leq \Theta \leq 1$, to estimate the individual variations in strength of links. Only links exceeding the threshold value Q^*=0.30 are included in final lexicons. It should be noted that the results remain nearly intact if threshold values lower than 0.30 are used because most below that threshold have very low values $Q < 0.2$, and thus, do not affect the correlation centrality in any essential way.

For the constructed individual lexicons, the importance of nodes and their rankings is based on correlation centrality as defined in Eq. (3). Lexicons then appear quite different, although it will be seen that many of them have about six to eight common terms, which have a high correlation centrality (i.e. key terms). Figure 2 shows eight lexicons that occur quite frequently as similar to other lexicons. It is obvious that it is difficult to compare the lexicons by visual inspection or to figure out how similar or different they are, or how key-terms appear in the set of 100 analyzed here. Nevertheless, Figs. 1 and 2 provide an overall idea of the appearance of the individual lexicons.

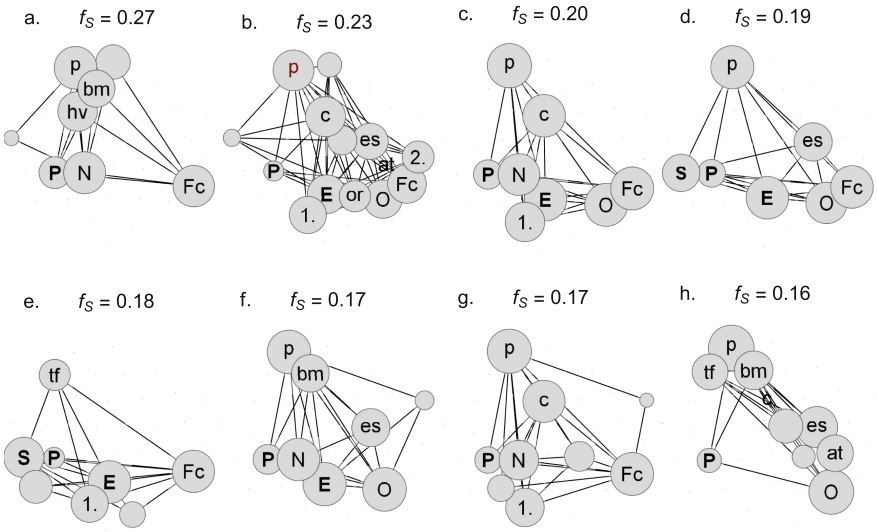

Fig. 2. Eight lexicons (from a to h) that appear most often in high similarity pairs, exceeding threshold $S^* = 0.55$. The frequency f_S of their appearance in similarity pairs is denoted.

To get an idea of importance of key-terms in the individual lexicons, we form rankings on the basis of correlation centrality Γ_k of a given term k in each individual lexicon, and then, find out how often the terms appear in the top-five, -ten and -15 cohorts of key-terms. Obviously, if a certain term appears in the top-five cohort in many individual lexicons, it is a key-term that deserves attention; the same applies, to a more moderate degree, to those terms that appear in the top-ten and -15 cohorts most often. Table 1 shows the rankings of key concepts, according to the different measures used as the basis for rankings.

The simplest measure is the frequency of occurrence (f) based ranking R_f. On the basis of the frequency of their appearance, the highest ranking concepts are, "plant" (occurs 208 times in 100 lexicons, corresponding to relative frequency f=1.0), "food chain" (f=0.45), "energy" (f=0.44), "producer" (f= 0.38) and "organism" (f= 0.28). These concepts are relevant in explaining the role of

Table 1. The key-terms and their rankings R_X appearing in students' responses. The subscripts X denote rankings based on: frequencies (f); top-15 terms for total range of phases $-1 \leq \Theta \leq 1$ as a stable reference value (r) and three cases of top-5 terms, which correspond to cohorts corresponding to total range (T), low quantile (L) and high quantile (H), as explained in the main text. Acronyms of terms in column A are as in Figs. 1 and 2

Term	A	R_f		R_r		R_T	R_L	R_H	Term	R_f	R_r	R_T	R_L	R_H
Plant	P	1	1.0	1	0.96	7	3	16	Decomposer	18	18	18	17	17
Food chain	Fc	2	0.45	3	0.70	2	2	2	Oxygen	19	22	20	23	22
Energy	E	3	0.44	4	0.56	3	6	3	Inorganic	20	23	23	20	25
Producer	P	4	0.38	2	0.71	1	1	1	Solar energy	21	20	16	14	20
Organism	O	5	0.28	5	0.47	4	5	4	Sugar	22	25	24	27	21
Nourishment	N	6	0.23	6	0.40	6	7	5	Photosynthesis	23	19	21	26	15
Consumer	C	7	0.18	7	0.34	5	4	6	Carbon dioxide	24	29	29	29	27
Ecosystem	Es	8	0.12	9	0.26	9	9	7	Carbon	25	–	–	–	–
Autotrophic	At	9	0.12	12	0.22	11	12	12	Nutrition	26	–	–	–	–
The first	1	10	0.11	8	0.27	8	8	8	Heterotrophic	27	24	19	15	23
Glucose	Gl	11	0.11	14	0.19	13	16	11	Food source	28	32	–	–	–
Biomass	Bm	12	0.10	11	0.22	14	19	9	Animal	29	26	27	21	–
Sun	S	13	0.10	10	0.24	12	11	13	Carnivore	30	–	–	–	–
Herbivore	Hv	14	0.09	13	0.20	10	10	10	Life	31	34	–	–	–
Organic	Or	15	0.08	17	0.14	17	18	14	Energy source	32	–	–	–	–
The second	2	16	0.07	15	0.15	15	13	19	Primary prod	33	27	–	–	–
Trophy level	Tf	17	0.06	16	0.14	22	24	18	Predator	–	31	28	28	-

plants as primary producers in the biosphere. In addition, concepts referring to systems, like "ecosystem", "the first" and "the second" (referring to levels in food chain) are common in the texts. However, for example, the occurrence of the concept of "photosynthesis" is relatively rare. Furthermore, concepts such as "the Sun" or "light energy", that refer to the origin of the energy, received less attention than expected.

The rankings of terms in Table 1 as based on values of correlation centrality as defined in Eq. (3) and calculated on basis of concurrence provide a picture somewhat different from frequency based rankings. The lowest value of concurrence Q obtained for $\Theta = 1$ is roughly proportional to classical contingency as measured by Yule's Y-factor multiplied by a factor Q_0 (see Appendix A) corresponding to correlations that are less affected by entanglement. The maximal co-occurrence with maximal entanglement is obtained with $\Theta = -1$. However, now we do not have enough information of individual lexicons, and thus, we generate ensembles of possible individual lexicons by letting parameter Θ vary. These ensembles of lexicons provide bounds for variance of diversity of connections but preserve the average frequency counts of corresponding to aggregated lexicons. With the ensemble of possible individual lexicons, we can estimate the ranking of terms based on correlation centrality by counting how often a given concept appears among the cohorts of top-five, -ten or -15 terms within the

ensembles Low (L), High (H) and Total (T). Rankings based on these ensembles are summarized in Table 1 for top-5 cohort denoted by R_X, with $X \in T, H, L$ and for top-15 cohort for ensemble T, which is a reference (denoted by subscript r) most closely related to frequency-based counting (i.e. containing low correlations due to co-occurrence) but selective about the top (15) key terms .

From the results in Table 1, we see that some key-terms have different rankings from the frequency-based rankings. For example, in ensemble H, the highest ranking term is now "producer" and the term "plant" has a much reduced ranking, down to R_{H}=16 from $R_{\mathrm{f}} = 1$. As expected, however, among an extensive enough cohort of top-15 terms ensemble, the term "plant" has rank 1 as in frequency-based ranking. This shows that while the term "plant" appears quite often, its entangled connections to other terms are weak, i.e. it appears in trivial ways in many sentences (e.g. as parts of lists, not in more complicated correlative connections). Such a conclusion is in agreement with results obtained previously, where "plant" is often connected to auxiliary other terms. Therefore, in the current analysis, with increased weight on entanglement, its importance and rankings are diminished. Due to entanglement based correlations, "producer" is in the topmost ranked term in the top-5 ensemble. This can be interpreted as "producer" being a kind of cohesion bringing term, connecting several other terms in a network through different sentences, not only through unstructured statements or lists. Some other terms, like "photosynthesis", "biomass", and "herbivore" also have increased rankings in comparison to frequency-based rankings. However, many terms retain their rankings as they appear in frequency-based analysis.

Finally, it is interesting to compare the similarity of lexicons. Figure 3 shows the distribution (in the form of a box-whisker plot) of normalized similarities S_N plotted against unnormalized similarities corresponding to values $S_0 = 1, 2, 3, \ldots, 7$ averaged (binned) over a range ± 0.5. The values of S_0 correspond roughly to a number of important nodes, which is at most about 6–7 in the most extensive individual lexicons. As seen from the box-whisker plots, values S_N and S_0 correlate, but S_0 is a better indicator for comparisons of similarities, because S_N has wide bounds of variation for most values of S_0. The box-whisker plots show cases obtained for cohorts Low (L), High (H) and Total (T). The corresponding histograms for probability density distributions p are shown in the lower row, in Fig. 3d–f. It is seen that Θ has a clear and significant effect on distributions of similarity values, high values of Q corresponding to substantial increased similarities. These results can be compared with Fig. 2 displaying eight examples of lexicons which appear most often in pairs of high similarity lexicons.

In the results shown in Fig. 3 the high values of Q (i.e. corresponding to increased correlations due to entanglement) are responsible of increased similarity of lexicons for cohort High (H) as seen in Fig 3c and in Fig 3f, where peaks in histogram of similarity distributions is due to such terms. This means that the terms with large values for correlation centrality are globally the most important, overall coherence-producing terms for all lexicons, and thus, the most important terms for learning language of science. In particular, term "producer"

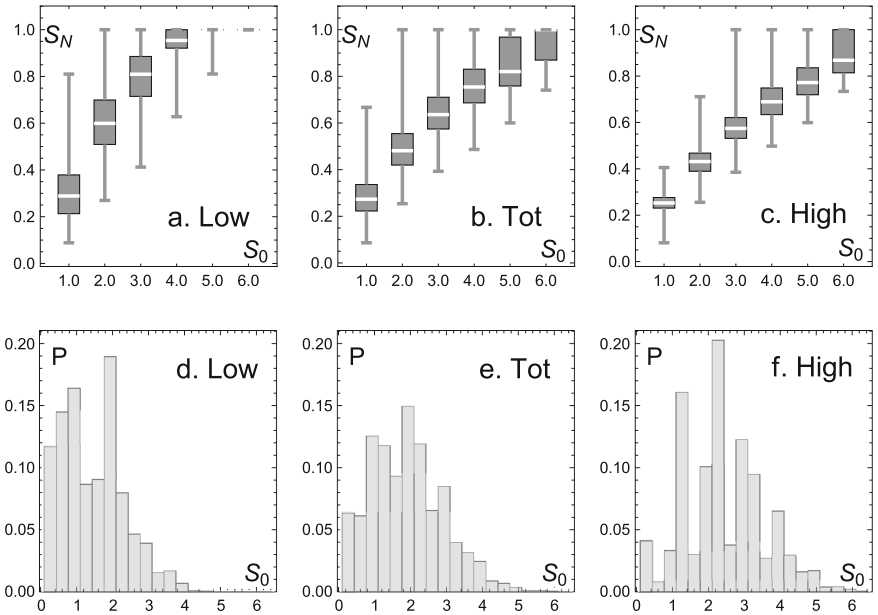

Fig. 3. Distributions of cosine- similarity values S_N and unnormalized similarity values S_0 for all lexicon pairs. The upper row shows a box-whisker plots of distribution of (normalized) cosine-similarities S_N when unnormalized values S_N, fall in the range $k - 0.5 < S_0 < k + 0.5$, with $k = 1, 2, 3, \ldots, 6$. Note that 99 % of values of S_0 are less than seven. Distributions are shown for cases corresponding to lowest (Low) and highest (High) quantiles of phase factors Θ, to be compared with values obtained for total range (Tot) of phase factors. The distributions are shown as standard box-whisker plots. Corresponding histograms of distributions of S_0 are shown in the lower row in the form of probability density p

attains then attains the most important role instead of "plant", which appears as the most important concept in frequency-based analysis as well as in traditional analysis ignoring entanglement (compare also results reported in ref. [5]). It is of importance that quantum sematics provides now a completely different picture of the importance of this ter. On the other hand, results obtained for low values of Q, where entanglement reduces correlations, do not differ significantly from frequency-based rankings nor from cases corresponding to total range of values (Tot), which minimize all correlations.

4 Discussion and Conclusions

The quantum semantics approach to constructing lexical networks (lexicons) that correspond to short written texts, with sparse co-occurrence of key terms, was applied to a sample of 100 students' texts about a topic in life -science (specifically, role of plants in food chains). The method allowed us to use entanglement

to model the effects of individual subjective bias in weighting the importance of co-occurrence of key terms in the text and thus, to use aggregated lexicons as a starting point to generate model ensembles of individual lexicons. Previously, with the same sample of texts, such analysis of individual lexicons was not possible due to the shortness of the texts, when co-occurrences of words in a given text are rare.

The results are mostly in line with previous analysis as well as with basic occurrence frequency counts. Nevertheless, the generalized co-occurrence counts and lexicons based on them showed some interesting and important differences. The key-terms that appear to be most central on the basis of their frequency of appearance turn out to not always be central in co-occurrence analysis. When the co-occurrence correlations correspond to maximal entanglement, the changes in rankings of key-terms can be significant. For example, the most frequent term ("plant") drops to a lower-ranking position and another term, not so frequently occurring ("producer") but more connected (entangled) in the lexical network takes the highest ranking position. The cohorts of top-five, -10 and -15 terms remain nearly the same but within the cohorts, changes in rankings may take place. The key-terms are identified as terms, which attain high values of concurrence and which are shared by many lexicons (i.e. lexicons having high similarity). Such terms are supposedly also the most important ones for learning the language of science.

The method introduced here is meant to be simple enough for practical applications in educational research, where advanced and complicated methods of text analysis are not likely to be widely adopted. However, it is useful to be able to go beyond frequency-counts and estimate co-occurrence of word-pairs and to construct lexical networks. The case study presented here suggests that the generalized method introduced here is suitable for such a purpose. The ability to estimate such variations is useful for educational applications, as a first step to give an idea of how the appearance of terms in a text may be related to different ways to read the texts.

Appendix A

In quantum semantics [12], word co-occurrence is described as a two-state event: words A and B may both occur at least once in a block of text, only one of them might occur (at least once), or neither A or B occurs. Occurrence is described by tag 1 while non-occurrence is described by 0. Co-occurrence can be then coded as pairs $|00\rangle$ (neither A nor B occur), $|01\rangle$ (A does not occur but B occurs), $|10\rangle$ (A occurs but not B), $|11\rangle$ (A and B both occur). These are the basic building blocks of the qubit states of two two-state systems [12,13]. The frequencies of each of them, found by counting co-occurrence, are denoted by n_{00}, n_{01}, n_{10} and n_{11}, respectively. We can construct a superposition of the two qubit states, corresponding to all potential ways to build up the word meaning as it is realized in different sentences [12]

$$|\Psi\rangle = c_{00}|00\rangle + c_{01}|01\rangle + c_{10}|10\rangle + c_{11}|11\rangle \tag{A.1}$$

where amplitudes $c_{ij} = \sqrt{n_{ij}}\, e^{i\phi ij}$ are complex numbers related to real valued frequency n_{ij} with normalization $\Sigma n_{ij} = \Sigma |c_{ij}|^2 = 1$ and also including phase factors ϕ_{ij}. The phase factors ϕ_{ij} can be taken as parameters that describe various subjective ways to read meanings into the co-occurrence of terms A and B, i.e. to emphasize the meaning of connection differently [12–14]. The way to model subjective bias through entanglement is a first step to overcome a situation, where co-occurrence frequency is completely fixed and thus, needs to be associated with only one possible way to contribute on the meaning of words through their occurrences in sentences.

Factorization of state $|\Psi\rangle$ in Eq. (A.1) into independent marginal states of one qubits is not possible in general; the state is said to be entangled. Different entanglements then correspond to different subjective, biased readings of the meaning of co-occurrence. The degree of entanglement can be quantified as a difference between the factorized state and the entangled state, called a concurrence Q. Here, we omit the details of the derivation (lengthy but straightforward) of concurrence, to be found elsewhere [12] and state the final result

$$Q = 2|c_{01}c_{10} - c_{00}c_{11}|\,, \quad 0 \leq Q \leq 1 \tag{A.2}$$

Substituting in Eq. (A.3) the amplitudes in Eq. (A.2) we finally get [12]

$$Q = 2\sqrt{n_{01}n_{10} + n_{00}n_{11}}\sqrt{(1 - \Theta\frac{\sqrt{n_{01}n_{10}n_{00}n_{11}}}{(n_{01}n_{10} + n_{00}n_{11})/2}}, \tag{A.3}$$

where $\Theta = \cos(\phi_{01} + \phi_{10} - \phi_{00} - \phi_{11})$ is the compound phase factor, taken here as the free parameter to account for (unknown) phase factors. The concurrence in Eq. A.2, using abbreviations Q_0 and R for the prefactor and fractional term, leads then to expression$Q = Q_0\sqrt{(1 - \Theta R)}$ of Eq. (1).

An analogy of concurrence to a so-called mean square contingency factor and Yule-factors [15,16] is interesting to note [12]. The mean square contingency (closely related to the so-called odds -ratio) is proportional to a factor $n_{11}n_{00} - n_{01}n_{01}$, to be compared with (A.2). The factor in Eq. A.2, without phase factors ϕ, would yield a factor $|\sqrt{n_{11}n_{00}} - \sqrt{n_{01}n_{01}}|$ reminiscent of Yule's Y-factor [16]. The main difference with concurrence Q and different forms of such contingency measures is that contingency measures are made independent of marginal distributions of co-occurrences, but here, it is important to preserve the total probability (rarely occurring co-occurrences are of no interest), taken into account by factor Q_0. Of most importance here is the notion that all measures of contingency as well as the concurrence Q take a value of zero when all clauses are randomly mixed [15,16] (see also ref. [12]), in which case associations between occurrence of words A and B disappear. The same condition also leads to $Q = 0$ and thus, to the disappearance of entanglement.

References

1. Fang, Z.: The language demands of science reading in middle school. Int. J. Sci. Educ. **28**, 491–520 (2006)
2. Ford, A., Peat, F.D.: The role of language in science. Found. Phys. **18**, 1233–1242 (1988)
3. Lemke, J.: Talking science: language. Learning and values. Language and educational processes. Ablex Publishing Corporation, Norwood, NJ (1990)
4. Yore, L.D., Hand, B., Goldman, S.R., Hilderbrand, G.M., Osborne, J.F., Treagust, D.F., Wallace, C.S.: New directions in language and science education research. Read. Res. Quart. **39**, 347–352 (2004)
5. Södervik, I., Nousiainen, M., Koponen, I.T.: First-year life science students' understanding of the role of plants in the ecosystem—a concept network analysis. Educ. Sci. **11**, 369 (2021)
6. Yun, E., Park, Y.: Extraction of scientific semantic networks from science textbooks and comparison with science teachers' spoken language by text network analysis. Int. J. Sci. Educ. **40**, 2118–2136 (2018)
7. Nousiainen, M., Koponen, I.T.: Pre-service teachers' declarative knowledge of wave-particle dualism of electrons and photons: finding lexicons by using network analysis. Educ. Sci. **10**, 76 (2020)
8. Christianson, N.H., Sizemore, B.A., Bassett, D.S.: Architecture and evolution of semantic networks in mathematics texts. Proc. R. Soc. A **476**, 20190741 (2020)
9. Chai, L.R., Zhou, D., Bassett, D.S.: Evolution of semantic networks in biomedical texts. J. Complex Netw. **8**, cnz023 (2020)
10. Ribeiro, E., Teixeira, A.S., Ribeiro, R., de Matos, D.M.: Semantic frame induction through the detection of communities of verbs and their arguments. Appl. Netw. Sci. **5**, 69 (2020)
11. Medeuov, D., Roth, C., Puzyreva, K., Basov, N.: Appraising discrepancies and similarities in semantic networks using concept-centered subnetworks. Appl. Netw. Sci. **6**, 66 (2021)
12. Surov, I.A., Semenenko, E., Platonov, A.V., Bessmertny, I.A., Galofaro, F., Toffano, Z., Khrennikov, A.Y., Alodjants, A.P.: Quantum semantics of text perception. Sci. Rep. **11**, 4193 (2021)
13. Surov, I.A.: Quantum cognitive triad: semantic geometry of context representation. Found. Sci. **26**, 947–975 (2021)
14. Galofaro, F., Toffano, Z., Doan, B.-L.: A quantum-based semiotic model for textual semantics. Kybernetes **47**, 307–320 (2018)
15. Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis. Pearson, New Jersey (2008)
16. Bonett, D.G., Price, R.M.: Statistical inference for generalized yule coefficients in 2x2 contingency tables. Soc. Meth. Res. **35**, 429–446 (2007)
17. Estrada, E.: Informational cost and networks navigability. Appl. Math. Comp. **397**, 125914 (2021)
18. Estrada, E.: The Structure of Complex Networks: Theory and Applications. Oxford University Press, Oxford (2012)
19. Estrada, E., Hatano, N., Benzi, M.: The physics of communicability in complex networks. Phys. Rep. **514**, 89–119 (2012)
20. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques, 3rd edn. Elsevier, Amsterdam (2012)
21. Newman, M.: Networks, 2nd edn. Oxford University Press, Oxford (2018)