Scientometrics (2022) 127:1755–1781
https://doi.org/10.1007/s11192-022-04287-9

The relationship between citations and the linguistic traits of specific academic discourse communities identified by using social network analysis

Don Watson[1] · **Manfred Krug**[1] · **Claus-Christian Carbon**[2]

Received: 22 February 2021 / Accepted: 27 January 2022 / Published online: 21 February 2022

Abstract
For a research article (RA) to be accepted, not only for publication, but also by its readers, it must display proficiency in the content, methodologies and discourse conventions of its specific discipline. While numerous studies have investigated the linguistic characteristics of different research disciplines, none have utilised Social Network Analysis techniques to identify communities prior to analysing their language use. This study aims to investigate the language use of three highly specific research communities in the fields of Psychology, Physics and Sports Medicine. We were interested in how these language features are related to the total number of citations, the eigencentrality within the community and the intra-network citations of the individual RAs. Applying Biber's Multidimensional Analysis approach, a total of 771 RA abstracts published between 2010 and 2019 were analysed. We evaluated correlations between one of three network characteristics (citations, eigencentrality and in-degree), the corpora's dimensions and 72 individual language features. The pattern of correlations suggest that features cited by other RAs within the discourse community network are in almost all cases different from those that are cited by RAs from outside the network. This finding highlights the challenges of writing for both a discipline-specific and a wider audience.

Keywords Specificity · Discourse community · Multidimensional analysis · Social Network Analysis · Research article · Communication · Dissemination

Introduction

The question of which factors contribute to the success of a research article (RA) has increasingly become an area of interest in the field of scientometrics (e.g. Barnett et al., 2011; Jamali & Nikzad, 2011; Lei & Yan, 2016; Nair & Gibbert, 2016; Chen et al., 2020; Colladon et al., 2020). Success is most often measured by citations or other metrics (such

✉ Don Watson
don.watson@uni-bamberg.de

[1] Department of English Linguistics, University of Bamberg, Bavaria, Germany

[2] Department of General Psychology and Methodology, University of Bamberg, Bavaria, Germany

as impact factor) that are derived from citations. Being cited by other scholars in one's discipline is a sign of acceptance, if not necessarily agreement, by the community of peers that make up the discipline. Using the concept of *discourse community* to describe the members of an academic discipline, we see that one defining feature of discourse communities is the way in which it uses language to communicate. Furthermore, each discipline uses language in its own very specific ways.

Academic communities communicate, not exclusively but most formally, through RAs. As each RA can be linked to others through citations, these communities have the characteristics of a social network i.e., each RA is a node in a network and each citation a link. Thus these communities can be investigated using social network analysis (SNA).

Using SNA to identify some very specific discourse communities within three distinct and diverse academic disciplines (Psychology, Physics and Sports Medicine), this paper will address two issues. First, it will attempt to describe the patterns of language use of these communities. Secondly, it will investigate if there are any relationships between the language use and the nature of their networks, in this case between frequency of grammatical items and either (a) the total number of citations received, (b) the number of citations received from other RAs within the network or (c) the centrality of RAs within their network.

Discourse communities

The idea of a discourse community, introduced by John Swales (1990), provides an understanding of how texts produce meaning through interaction and how authors' linguistic choices depend on purposes, context and audience. Swales gave six defining characteristics that discourse communities possess:

1. broadly agreed set of common public goals,
2. mechanisms of intercommunication among its members,
3. participatory mechanisms used primarily to provide information and feedback,
4. utilisation and possession of one or more genres in the communicative furtherance of its aims,
5. some specific lexis,
6. a threshold level of members with a suitable degree of relevant content and discoursal expertise (Swales, 1990, pp. 24–27)

Characteristics two, four and five all point to the central role of language use in the definition of a discourse community. Research writing, of course, is not homogeneous but consists of a variety of specific, albeit related, genres that are used by different discourse communities. This concept of *Specificity* has emerged because "many language features, including vocabulary, are specific to particular disciplines" (Hyland & Tse, 2007, p. 251). Thus any investigation into the nature of language use in academic research needs to focus on specific disciplines individually as there are likely to be differences between research disciplines.

Studies that have considered this topic have typically involved genre analyses of RAs and focused on moves (e.g., Ebrahimi & Chan, 2015; Kim, 2014; Pho, 2008; Tseng, 2011) or lexico-grammatical features. Kim (2014), for instance, looked at move patterns in abstracts in the social sciences, Pho (2008) investigated the rhetorical moves of abstracts in the fields of Applied Linguistics and Educational Technology and the linguistic realizations

of moves and authorial stance in different abstract moves. Ebrahimi and Chan (2015) analysed and compared the discourse functions of grammatical subjects used in RA abstracts in the disciplines of Applied Linguistics and Economics, and Tseng (2011) analysed move structure and verb tense in Applied Linguistics abstracts.

The methodologies of the studies mentioned, however, have not considered the specificity of the discourse communities from which the RAs came. Previous research has generally taken a random or more often semi-random selection of RAs from popular journals in various fields as their corpus. For instance, Tseng (2011) took 90 RA abstracts from three of the top seven journals in Applied Linguistics as they represented the, "status quo" (Tseng, 2011, p. 29) in the field. Although this makes an attempt at considering specificity, even Applied Linguistics has multiple foci of research and differing methodologies. The present study will attempt to address this limitation by using a shared topic being investigated by the RAs as the initial selection criteria for the corpus.

Apart from a shared focus on a particular research topic, there is of course another aspect, unique to RAs, which can aid in the identification of specific research communities. Citations offer an explicit and measurable set of relationships between RAs that can be seen to be exhibits of points two and three of Swales' defining characteristics of discourse communities. But more than simply offering mechanisms of intercommunication and participation, authors are able to make their claim to a position within a community visible by citing others (Hewings, Lillas & Vladimirou, 2010). In addition to this, one could consider the accumulation of citations as an indicator of the preferences of the members of the discourse community. Therefore, through considering citations, we can begin to see how the "participatory mechanisms used primarily to provide information and feedback" (i.e. Swales' third point above) within a discourse community could be measured.

Citations and linguistic measures

In the field of Scientometrics, numerous studies have used citation data to investigate the relationships between various linguistic traits (such as titles, for example) and non-linguistic traits and citation rates. Several of these studies shared a previously discussed limitation, in that their corpora were chosen from an arbitrary set of journals. Lei and Yan (2016 analysed the readability of abstracts and full texts of RAs in the field of Information Science and investigated whether readability scores were correlated with the number of citations. The study limited its corpus selection to four journals; these were chosen as they were considered, "important journals in information science" (Lei & Yan, 2016, p. 1157). Similarly, Dolnicar and Chapple (2015) studied the association between readability and citations in tourism journals and compiled a corpus from a small selection of three journals with the highest impact factors.

Other studies have different limitations. Gazni (2011) investigated if the readability of RA abstracts correlated with their scientific impact, however, the data was collected across numerous academic fields, and thus no discourse community-specific conclusions could be drawn. Nair and Gibbert (2016), Jacques and Sebire (2010), Jamali and Nikzad (2011), Paiva, Lima and Paiva (2012), and Subotic and Mukherjee (2014) all looked at the attributes of RA titles and their relationship with citation counts. However, the brevity of titles means that little information can be gleaned as to how the fields studied use language more generally. Hartley, Sotto and Pennebaker (2002) found that highly influential RAs were more readable than less influential ones. Rather than any actual measure of citations, this study used a list of the 100 most influential journal articles in cognitive science in the

twentieth century posted on the World Wide Web (WWW) as their corpus of highly influential RAs. This was then compared to a control corpus of RAs taken from the same journal editions as the highly influential ones. The limitation of this study is that the wealth of data that exists in moderately cited RAs is excluded and thus a full picture of the discourse community as a whole and its relationship with its language use remains underexplored.

Similarly, three further studies have compared linguistic traits of RAs grouped according to citations or impact. Jin (2018) compared two corpora of discussion sections in the field of Chemical Engineering using a Multidimensional Analysis (MDA) approach. The articles were grouped into a "Corpus of High-impact" RAs taken from highly cited articles from high impact journals and a "Corpus of Low-impact" RAs with few citations from, "less recognized peripheral journals". Lu et al. (2019) selected RAs from the Public Library of Science (PLoS) from the fields of Biology and Psychology. Their analysis considered 12 variables of linguistic complexity: sentence length per article, standard deviation of sentence length per article, type-token ratio, clause ratio, and the length and ratio for nouns, verbs, adjectives and adverbs. They categorised the articles into three groups: high impact (top 1% most cited papers), medium impact top 10% without the top 1%), and low impact (the remaining 90%). Chen et al. (2020) selected RAs from PLoS in Biology, Genetics and Medicine-related fields. Their analysis used seven indicators of linguistic characteristics (title length, abstract length, full-text length, sentence length, lexical diversity, lexical density and lexical sophistication) and categorised the articles into the top 20%, bottom 20% and total of the viewed and downloaded articles. Of these three studies, only Jin (2018) found any relationship between linguistic traits and citations or impact. Their analyses suggested that more "expert" performances incorporated more metadiscursive features, first-person pronouns and evaluative statements with further explanation. However, all three studies grouped the RAs into arbitrary levels of citations/impact and analysed their data for differences between these groups.

Although all of these studies used citations as variables in their analyses, none of these studies considered if the corpora represented genuine communities by exploring if and how RAs were citing each other and how this influenced language use. The question of how to understand the citation ties between RAs is where we turn next.

Social network analysis

Social Network Analysis (SNA) has become an accepted method for analysing a broad variety of phenomena that can be conceptualised as nodes and the edges or ties that connect them. Various measures of the characteristics of networks have been derived such as the clustering coefficient, which measures the prevalence of cliques or smaller highly interconnected groups within a network. Average path-length measures the average distance, measured by the connections between nodes, between all the nodes in a network. The most well-known of these characteristics is probably smallworldness. This has been shown to be a feature of most real-world networks (Humphries & Gurney, 2008). Being small world means that the network is neither regular (i.e. a uniform lattice) nor random, but somewhere in between. Smallworldness, then, is characterised by a high clustering coefficient and a short average path length. Another useful measure is eigencentrality, which is a measure of the relative influence a node has in a network (Spizzirri, 2011).

Owing to this variety of insightful measures, SNA has become a popular tool to study various attributes of academic publications. These include the social capital of authors (Jha & Welch, 2010; Nahapiet & Ghoshal, 1998), mapping the structure of publication

networks, (Barnett et al., 2011; Behara et al., 2014; Chen, Baird & Straub, 2014; Agnoloni, 2014) and analysing the content of texts (Galvez, 2019; Busse, Gather & Kleiber, 2016). Few studies have explored the relationship between the language used in RAs and the social networks that they are part of. An exception is Colladon et al. (2020). This study attempted to use social network and semantic analysis to predict the future success of RAs using publications in the field of Chemical Engineering. The semantic features considered were abstract length, sentiment, complexity, lexical diversity and commonness. Although they found strong correlations between other variables and citations, only moderate correlations were found between citations and the semantic variables. They concluded that writing longer, more informative abstracts somewhat contribute to publication success. However, similar to all the previously mentioned studies, the corpus was not collected by selecting RAs that cited each other and were therefore demonstrably in the same discourse community; rather, they were selected using RAs with the same All Science Journal Classification (ASJC) tag. Furthermore, the network variables used were based on the author network, i.e. the nodes in the network were individual authors. This last point should not be considered a limitation, however, but simply a difference between that study and the present one.

Linguistic measures

The studies which have investigated language use discussed thus far have considered a wide variety of linguistic measures that each have their own merits. This paper will employ Biber's (1988) Multidimensional Analysis (MDA) technique, which explores genre variation using large text corpora and statistical tools, most notably factor analysis. The first reason for this approach to the present data is that, unlike measures such as readability or complexity, MDA does not measure theoretical constructs that are the subject of ongoing debates (see e.g., Begeny & Greene, 2014; Kortmann & Szmrecsanyi, 2012). MDA merely uses computational tools to tag words in texts for their lexical and morpho-syntactic categories. Frequency counts of linguistic features are then carried out within texts, and the distributions compared across texts (Biber, 1992). MDA is generally used to identify co-occurring distributions of linguistic features, which are rendered as numbered "dimensions". These dimensions are then interpreted and given a descriptive label. This interpretation and labelling is the principal input of the researcher. The second important reason for our approach is that conducting a MDA yields results on two levels of analysis: the primary level is constituted by the frequency distributions of all individual lexical and morpho-syntactic features; the secondary level consists of the identified dimensions, each of which includes a selection of the relevant primary features and their loadings, i.e. their relative contribution to the respective dimension. This study will exploit both levels.

Research question

The present study has the following two research questions. Firstly, which lexical, grammatical and semantic features co-occur in RA abstracts published on three specific research topics in the fields of Physics, Psychology and Sports Science? Secondly, are these language features correlated with the characteristics of each of the citation networks including total citations, eigencentrality and in-degree (i.e. the number of citations received only from within the citation network)?

Method

Corpora selection

The corpora for this study were compiled from the Web of Science Core Collection. Three separate search terms were used initially. These were "post-traumatic stress disorder" or "PTSD" (the commonly used abbreviation), "Higgs Boson" and "Endurance training". The three searches were then filtered by publication year (2010–2019) and document type (article). This gave 8129 results for "post-traumatic stress disorder" (PTSD), 4897 results for "Higgs Boson" (HB) and 1722 results for "endurance training" (ET). The results were then exported as "full record and cited references".

Network analysis

The exported files were analysed with the Network Analysis Interface for Literature Studies (NAILS cf. Knutas et al., 2015) in order to produce node and edge files of the bibliometric network. These files could then be imported into the open source network visualisation software Gephi (Bastian, Heymann & Jacomy, 2009), which was then used to filter the network to identify the specific research communities for the analysis. The network was first filtered using the "giant component query". A component in network theory is a group of nodes (in this case RAs) that are connected to each other. The "giant component" filter identifies the largest of these components in the network and excludes all others. This means that all nodes in the network are connected, either directly or via other nodes. The network was then sub-filtered using both "in-degree range" (citations received) and "out-degree range" (citations given) with the range parameters set to a minimum of one. This method ensures that all RAs were in one contiguous network and part of the same community of communicating researchers. Gephi was further used to calculate the eigencentrality and in-degree of all RAs in the subsequent networks, as well as the clustering coefficient and average path length of each of the three networks. The smallworldness index was calculated for each network using the *qgraph* package (Epskamp et al., 2012) in R 4.0.4 (R Core Team, 2021).

Multidimensional analysis

Once the specific communities of RAs had been identified, a multidimensional analysis (MDA) of linguistic variables for each of the corpora was conducted in order to extract linguistic dimensions. After removing any extraneous text such as copyright information, the abstracts were grammatically annotated, or tagged, using the open source software Multidimensional Analysis Tagger (MAT) (v. 1.2; Nini, 2014). Following the method suggested by Biber (1992), an exploratory factor analysis (EFA) was conducted on the data obtained from the MAT tagger to extract factors. As there were many tagged variables with very low scores (including many zeros), any variables with a mean of less than 0.2 per 1000 words were removed from the data before the factor analysis. R 4.0.4 was used to perform the EFA using the *psych* package (Revelle & Revelle, 2015). Several criteria were used to determine the best number of factors. These criteria included a visual inspection of the scree plot, the deflection point of the eigenvalues, the Tucker-Lewis index of factoring reliability and the interpretability of the resulting factors. After factor extraction, "Varimax" factor rotation was used to force each linguistic feature to load on as few factors as possible.

Only features with a loading of 0.35 or higher (following the method of Biber, 1992) were included in the factors and for any features that loaded on more than one dimension, only the highest loading was retained. The resulting dimensions list the co-occuring linguistic features and a weighting ranging from −1 to 1. By multiplying the *z*-score of the frequency of the linguistic features on a dimension by the weighting and calculating the mean of these weighted scores, a mean dimension score can be calculated for each RA.

Statistical analysis

In order to identify any correlations between linguistic variables and network variables, a correlation matrix of Pearson's *r* coefficients for all possible pairs was computed using the rcorr function in the *Hmisc* package (Harrell Jr, F.E. & Harrell Jr, M.F.E., 2019) in R. In the first such analysis, the matrix contained the mean dimension scores for each dimension for each RA, the total citations received by each RA, the in-degree or number of citations received by each RA from other members of its research community, and the eigencentrality score of each RA. In the second analysis, the matrix contained the *z*-scores of all tagged linguistic variables, the total citations received by each RA, the in-degree or number of citations received by each RA from other members of its research community and the eigencentrality score of each RA.

Results

Network statistics

Table 1 shows the descriptive statistics of the citation networks for each of the three corpora used in this study. As described in the Method section, each corpus was derived from an initial keyword search on Web of Science and then filtered using the network visualization software Gephi (Bastian, Heymann & Jacomy, 2009). The filtering process

Table 1 Descriptive statistics of RA citation networks

Corpora	Endurance training (ET)	Higgs Boson (HB)	Post-traumatic stress disorder (PTSD)
Downloaded references	1722	4897	8129
References after filtering	94%	501	176
% of total downloaded	5.4%	10.2%	2.2%
Total words	25,236	78,709	36,225
Mean citations	58.26	55.74	77.02
Range of citations	1–342	1–510	4–614
In-degree range	0–6	0–32	0–15
% in-degree cites	2.37	5.51	1.66
Clustering coefficient	0.072	0.116	0.039
Average path length	1.626	2.944	1.568
Smallworldness	12.887	6.494	24.037

had the largest effect on the PTSD corpus with only 2.2% of RAs remaining and the least on the HB corpus, which retained 10.2% of the total downloaded references. The RAs in the PTSD corpus were the most cited with a mean of 77.02 citations per RA whereas ET and HB showed similar numbers of citations with 58.26 and 55.74, respectively. Although the PTSD RAs had the most citations in total, the HB RAs had the most citations from other RAs within the corpora's citation network. The HB network received 5.51% of the total citations from within the corpora network compared to 2.37 and 1.66% for the ET and PTSD corpora, respectively. Similarly the HB network range of in-degree citations was higher than those of the other two networks. This should be expected given that the giant component filtering had the least effect on the HB network. It shows the HB network has the most citation inter-connections between RAs of the three and is thus the most self-contained of the networks. The PTSD network, by contrast, is the least self-contained and the one most connected to the wider academic citation network, with the ET network falling somewhere between the other two on these measures.

Regarding the network analysis statistics, the HB network showed the highest clustering coefficient and the highest average path length, indicating it was the least random network, while the PTSD showed a lower clustering coefficient and shorter average path length, indicating it was the most random. These figures are reflected in the smallworldness indices, where the PTSD network has the highest score and HB the lowest. Nonetheless, all three networks easily pass the accepted level of >1 (ET = 12.887; HB = 6.494; PTSD = 24.037) as an indicator that the network is a smallworld network and indeed they also pass the more stringent level suggested by Humphries and Gurney (2008) of >3. Figures 1, 2 and 3 show graphical representations of the networks. Nodes represent individual texts and the edges represent a citation link. The size of each node indicates the in-degree citation score or, in other words, the citations the text received from other texts within this network. The shading represents the total (intra- and extra-network) citations which that text received, with darker shades indicating more citations.

Fig. 1 "Endurance training" corpus: citation network. n.b. node size = in-degree citation score; node shade = number of citations (darker is more)

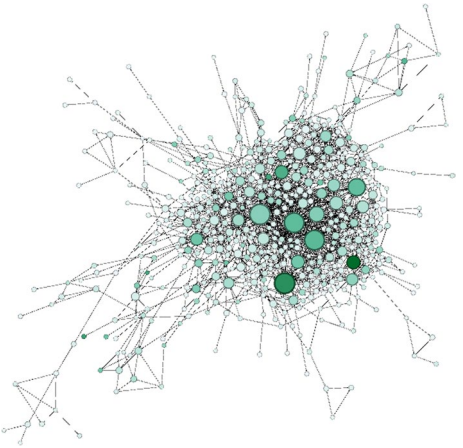

Fig. 2 "Higgs boson" corpus: citation network. n.b. node size = in-degree citation score; node shade = number of citations (darker is more)

Fig. 3 "Post-traumatic stress disorder" corpus: citation network. n.b. node size = in-degree citation score; node shade = number of citations (darker is more)

Multidimensional analysis results

Appendices 1–3 show the multi-dimensional analysis loadings of grammatical features for each of the three corpora. The first corpus, based on the search term "endurance training", was reduced to three dimensions. These are labelled "non-numerical conclusion statements", "indicating results" and "describing procedure". The second corpus based on the search term "higgs boson", was reduced to five dimensions. These are labelled "mathematical terminology", "copular constructions", "indicating tentative results" and "subordination strategies". The third corpus based on the search term "post-traumatic stress disorder", was reduced to six dimensions. These are labelled "present relevance", "comparison of actions, events or states", "non-numerical description", "subordinating strategies", "indicating tentative results" and "passive constructions".

Comparison between dimensions

Although not identical, the various dimensions on each of the corpora show significant overlaps and similarities. For instance, both HB and PTSD demonstrate a significant use of passive constructions and subordination strategies. However it is on the use of mathematical notation, cardinal numbers and symbols, where HB and PTSD show a different pattern. HB shows a tendency towards using mathematical and other notations (and highly specialized words tagged as 'foreign words' by the corpus tagger, e.g. phi, mu and tau), whereas ET and PTSD are characterized by dimensions with tendencies towards a lack of this kind of mathematical notation. Only some form of "indicating results" is present in all three corpora, with the difference being that the results are stated less tentatively (i.e. without the use of possibility modals such *may, can* or *could*) in the ET corpus.

Table 2 shows for each corpus the correlations between the mean dimension score of each individual RA and either the total citations that RA received, the eigencentrality score of the RA within its network (as described above) or the in-degree (citations received from other RAs within the network) of that RA. A maximum p value of ≤ 0.1 was used to indicate statistical significance. This was done in order to create better comparability with the results of Colladon et al. (2020) which is the most similar study to the present one. On the ET corpus, there were two statistically significant correlations. They were between total citations and Dimension 3 "describing procedure" ($r = 0.29$, $p = 0.0048$), and between in-degree and Dimension 1 "non-numerical conclusion statements" ($r = –0.23$, $p = 0.0280$). On the HB corpus, there was one statistically significant correlation between citations and Dimension 1 "mathematical terminology" ($r = 0.08$, $p = 0.0628$). On the PTSD corpus, there was one statistically significant correlation between total citations and Dimension 6 "passive constructions" ($r = 0.24$, $p = 0.0011$).

Correlations between individual grammatical features and network statistics

Table 3 shows the statistically significant correlations between individual grammatical features and citations, eigencentrality and in−degree for each corpus. The complete correlation matrices for all grammatical features tagged are shown in appendices 4–6. Each corpus was tagged for 79 different grammatical features. Thus for each corpus, there were 237 correlations calculated for each corpus (3 network measures × 79 grammatical measures). The ET corpus showed 24 correlations that reached at least a statistical significance level of $p = \leq 0.1$. Of these, the strongest correlations were with total citations. In the HB corpus, there were 32 correlations that reached at least a statistical significance level of $p = \leq 0.1$. In the PTSD corpus, there were 16 correlations that reached at least a statistical significance level of $p \leq 0.1$. In this corpus, unlike the others, the strongest of the correlations was with in-degree. In fact, the two strongest correlations in this corpus were both with place adverbials (with in-degree and eigencentrality). The ET and PTSD corpora showed similar patterns of having the highest number of statistically significant correlations with citations (11 and 6) and the least with in-degree (4 and 4), whereas in the HB corpus, the significant correlations were evenly spread across the different network measures (11 each for in-degree and eigencentrality and 10 for citations). This pattern whereby the HB corpus shows more emphasis on in-degree and eigencentrality than overall citations echoes the earlier finding that the HB network is the most self-contained (cf. Figs. 1, 2 and 3).

Table 2 Spearman's correlation coefficients (r) of network statistics and mean dimension scores

	Dimension 1 Non-numerical conclusion statement	Dimension 2 Indicating results	Dimension 3 Describing procedure
"Endurance Training" corpus			
Citations	−0.03	− 0.08	0.29**
Eigencentrality	−0.17	− 0.12	0.12
In-degree	−0.23**	− 0.15	− 0.02
$n = 94$			

	Dimension 1 Mathematical Terminology	Dimension 2 Copular Constructions	Dimension 3 Constructions	Dimension 4 Indicating Tentative Results	Dimension 5 Subordination Strategies
"Higgs Boson" corpus					
Citations	0.08*	−0.06	− 0.05	0.01	0.04
Eigencentrality	− 0.07	− 0.02	0.02	0.05	0.01
In-degree	− 0.02	0	0	0.06	0.01
$n = 501$					

	Dimension 1 Present Relevance	Dimension 2 Comparison of Actions, Events or States	Dimension 3 Non−numerical Description	Dimension 4 Subordination Strategies	Dimension 5 Indicating Tentative Results	Dimension 6 Passive Constructions
"Post− traumatic stress disorder" corpus						
Citations	−0.05	− 0.1	− 0.06	0.04	− 0.07	0.24**
Eigencentrality	0.05	− 0.07	0.1	− 0.01	−0.04	0.08
In-degree	0.06	− 0.05	0	− 0.04	− 0.02	0.012
$n = 176$						

Statistical significance: $^{**}p \leq 0.05$; $^{*}p \leq 0.1$

Discussion

This study has investigated the links between characteristics of specific academic discourse communities and their use of language features. Unlike previous studies, the present study compiled corpora with the aid of SNA techniques that resulted in corpora in which all the RAs used were from a contiguous citation network. This ensured that the corpora reflected a unique discourse community. The three compiled citation networks, each discussing a particular topic within their discipline, were analysed in terms of various network features such as the number of citations, eigencentrality within the network, citations from within the network (i.e. in-degree), clustering coefficient, average path length and smallworldness. The RAs were also analysed for their use of a wide variety of linguistic features. Furthermore, these language features were analysed to discover their co-occurring patterns of use. Finally, correlations were calculated to investigate if any relationship existed between the relevant network features and language use. The three RA networks displayed

Table 3 Spearman's correlation coefficients (r) of network statistics and statistically significant linguistic features

	Tokens	TTR	COND	DT	GER	IN
"Endurance Training" corpus						
Citations	0.34***	0.44***	− 0.07	− 0.18*	−0.26**	0.38***
Eigencentrality	0.12	0.24	− 0.1	− 0.17	0.05	0.15
In-degree	0.13	0.13	− 0.18*	−0.08	0.05	0.16
	PHC	PIT	POS	TO	TSUB	VB
Citations	− 0.17*	− 0.07	− 0.22*	− 0.13	0.47***	− 0.04
Eigencentrality	− 0.19*	− 0.09	− 0.12	−0.22**	0.3**	− 0.18*
In-degree	− 0.08	− 0.18*	− 0.17	−0.21	0.15	−0.24
	PASS	PEAS	PUBV	SERE	SMP	SPIN
citations	−0.29**	0.01	0.25**	0.02	− 0.15	0.18*
eigencentrality	−0.26**	−0.11	0.3**	0.17*	−0.1	0.02
In-degree	− 0.16	− 0.18*	0.14	0.13	−0.18*	− 0.08
$n = 94$						
	ANDC	DC	CONC	COND	DPAR	FPPI
"Higgs boson" corpus						
Citations	0.11**	− 0.08*	0.0.	0.03	0.18***	0.06
Eigencentrality	−0.05	− 0.03	0.15***	0.05	0	− 0.08*
In-degree	0.03	− 0.01	0.07	0.09**	− 0.01	−0.08*
	INPR	JJ	NEMD	NN	NOMZ	POMD
Citations	0.08*	0.03**	− 0.02	− 0.15***	0.07	−0.03
Eigencentrality	0.08*	0.05	− 0.02	−0.02	− 0.05	0.09**
In-degree	0.06	0.03	0.08*	− 0.02	−0.08*	0.07
	PRMD	PRP	RB	TO	TPP3	VBD
Citations	0.03	0.01	0.06	0.04	0.03	0.11**
Eigencentrality	0.12**	0.08*	0.05	− 0.09**	0.12**	− 0.02
In-degree	0.13**	0.04	0.08*	− 0.08**	0.08*	− 0.04
	VPRT	PEAS	SMP	SPAU	STPR	WHCL
Citations	0.06	0.03	− 0.06	0.1**	0.1**	0.09*
Eigencentrality	− 0.11**	0.09**	− 0.08*	0.04	− 0.05	0.05
In-degree	− 0.08	0.08*	− 0.07	0.08*	− 0.03	0.04
$n = 501$						
	Tokens	TTR	DEMO	DPAR	DWNT	JJ
"Post− traumatic stress disorder" corpus						
Citations	0.04	− 0.01	− 0.13*	0.21**	0.16**	− 0.17**
Eigencentrality	− 0.15**	− 0.13*	0	0	− 0.06	− 0.07
In-degree	− 0.04	− 0.05	− 0.1	0.03	− 0.04	− 0.13**

Table 3 (continued)

	LS	NN	NOMZ	PHC	PLACE	POMD
Citations	0.27***	0.1	0.21**	0.1	0.08	− 0.11
Eigencentrality	− 0.02	0.11	− 0.04	− 0.15**	0.33***	− 0.1
In-degree	0.11	0.19**	− 0.03	− 0.06	0.37***	− 0.14*
		PRMD		WZPRES		
Citations		0.04		− 0.02		
Eigencentrality		0.19**		0.16**		
In-degree		− 0.04		0.06		

$n = 176$ Statistical significance: *** $p \leq 0.001$; ** $p \leq 0.05$; * $p \leq 0.1$

ANDC Independent clause coordination; *CD* cardinal number; *CONC* concessive adverbial subordinator; *COND* conditional adverbial subordinator; *DEMO* demonstrative; *DPAR* discourse particle; *DT* determiner; *DWNT* downtoner; *FPP1* first person pronoun; *GER* gerund; *IN* preposition/subordinating conjunction; *INPR* indefinite pronoun; *JJ* attributive adjective; *LS* list item marker; *NEMD* necessity modal; *NN* other noun; *NOMZ* nominalisation; *PASS* agentless passive; *PEAS* perfect aspect; *PHC* phrasal coordination; *PIT* pronoun it; *PLACE* place adverbial; *POMD* possibility modal; *POS* possessive ending; *PRMD* predictive modal; *PRP* personal pronoun; *PUBV* public verb; *RB* adverb; *SERE* sentence relative; *SMP* seem/appear; *SPIN* split infinitive; *TO* infinitive; *TPP3* third person pronoun; *TSUB* that relative clauses on subject position; *TTR* type token ratio; *VB* verb, base form; *VBD* verb, past tense; *VPRT* present tense; *PEAS* perfect aspect; *SMP* seem/appear; *SPAU* split auxiliary; *STPR* stranded preposition; *WHCL* wh-clause; *XX0* analytic negation

a number of striking differences. Most notable, perhaps, was the percentage of citations that were received from within the network. The network of the corpus based on the search term "higgs boson" (HB) received considerably more intra-network citations compared to the other two: nearly double that of the network based on the search term "endurance training" (ET), and almost five times as many as the network based on the search term "post-traumatic stress disorder" (PTSD). It would seem from this statistic that this specific research area is relatively self-contained and represents the clearest example of a close-knit discourse community of the three considered in the present study. This interpretation is supported by the Web of Science subject categories to which each RA was assigned. The 176 PTSD RAs were assigned to 12 different subject categories, whereas the 501 HB RAs were assigned to only three. Furthermore, the least common of these three HB categories consisted of only three RAs. On all of these measures, the ET network fell somewhere in between.

As expected (see previous research by, e.g., Huang, 2018; McGrath & Kuteeva, 2012; Jiang & Hyland, 2018), the corpora, coming from such fundamentally different disciplines, displayed some notable differences in their use of language features. This should not be surprising given the differences in the methods used and objects of study of the different disciplines. The differences are especially noticeable in the dimensions that emphasise or de-emphasise the use of mathematical notation such as numbers, symbols and the words that represent them. It is, however, interesting to note the many overlaps between the dimensions against the backdrop of the debate within applied linguistics as to whether there is indeed an "academic" register or if this is too broad a term. The observed overlaps give a clearer idea of the patterns of language use that are widely enough used to be considered "academic" rather than specific to any discipline.

Arguably, however, it is the differences in language use between the corpora rather than the similarities that are most relevant. Here the correlations between language features and network characteristics can further highlight the subtle differences between disciplines regarding certain linguistic choices. For instance, the use of passive was positively correlated with eigencentrality in the PTSD corpus but was negatively correlated with citations and eigencentrality in the ET corpus. In the HB corpus, this correlation was close to zero (–0.03 and –0.02, respectively). Another example is attributive adjectives, which are positively correlated with citations in the HB data but negatively correlated with citations in the PTSD data. Nevertheless, attributive adjectives are slightly more common in the PTSD data (117 instances per 1000 words, compared with 108 in the Higgs Boson corpus) despite this being a less successful trait in the PTSD community. This indicates that the patterns of preference and dispreference for language features is complex and dependent on the particular feature.

It is important to note that the correlations identified in the present study are modest. Our findings do, however, parallel those of Colladon et al. (2020), who also found similar (and equally modest) correlations between citations and certain semantic features of abstracts in a comparable study. Modest correlations are what should be expected though, given that there are many factors that play a role in the success of a RA. Indeed many such factors have been well researched (e.g. Didegah & Thelwall, 2013; Jacques & Sebire, 2010; Jamali and Nikzad, 2011). The aim of this study was to investigate a hitherto neglected aspect of RAs that is nonetheless likely to play a role in publication success.

Modest correlations notwithstanding, there are some valuable insights to be gained from closely looking at the way in which the correlations are distributed. Although a multidimensional analysis is a powerful tool to identify patterns of language use across multiple linguistic features, correlations between network characteristics and the linguistic measures used in this study were more common for individual language features than for the dimensions. This is likely due to dimensions including multiple features, some of which did not correlate with a network characteristic. Those dimensions that did correlate contained individual features with high loadings that themselves correlated with network characteristics. A little over half of the significant correlations were with features that were a part of a dimension, and the more important they were within the dimension, the more likely that dimension was to be significantly correlated with a network characteristic. Perhaps it is the individual linguistic features that correlate with network characteristics, but are not part of one of the identified dimensions, which are especially characteristic. Certainly it is the less obvious non-dimensional features that would-be authors could profitably assimilate into their own writing in order to give themselves an edge.

Another intriguing aspect of the correlation pattern is that, with only two exceptions, linguistic features correlate either with total citations or in-degree, but not both. The exceptions were split auxiliaries in the HB corpus and attributive adjectives in the PTSD corpus, which clearly suggests that language features that are popular within the community are different to those that are popular with those that are outside of the community. A follow-up correlation analysis was conducted in which non-network citations were calculated by subtracting the in-degree number from the total citations for each RA. Correlations between linguistic features and non-network citations yielded only one additional statistically significant result for nominalisations on the HB corpus, which resulted in the correlation between nominalisations and intra-network citations (i.e. in-degree) being negative ($r = -0.08$, $p = 0.0644$), and between nominalisations and the non-network citations being positive ($r = 0.08$, $p = 0.0837$). This means that the use of nominalisations in RAs that discuss the topic "Higgs Boson" is associated with fewer citations from other RAs that

are within the immediate research community and more citations from RAs that are from outside it, which lends further support to the above conclusion. Moreover the observation is fully in line with (and qualifies further) the notion of linguistic specificity, which proposes that different discourse communities use language in their own specific ways. In this case, the very specific discourse communities under investigation in the present study have their own unique discoursal expectations and conventions, whereas those outside of it have a variety of differing expectations that may or may not overlap. Thus RA authors seem to have two different audiences with differing expectations. Although specific advice on how a RA author can deal with these dual audiences is beyond the scope of this paper, authors would be advised to consider their aims for, and expectations of, their RAs within their research community and beyond when drafting their papers.

Regarding the limitations of this study, linguistic features are, as previously mentioned, only one of many aspects of a RA that are related to citations and centrality. Nonetheless, with the strong imperative for academics to publish and, as established in the discussion of discourse communities, the need for authors to use discipline-specific language in order to be accepted by their peers, linguistic conventions represent one aspect that cannot be ignored. This focus on discipline specificity leads to another potential criticism of the present study, viz. that the corpora are *very* specific. However, it emerges exactly from the clear differences between the corpora under investigation in this study that any attempt to study the language of research *needs* to be *very* specific. In order to mitigate this potential issue, and to enhance the generalisability of its findings, this study has investigated a *broad selection* of very specific communities. From this follows of course that there are many other communities which may show other relationships to network features.

Conclusion

This study provides a unique perspective on the question of language use in academic discourse by using SNA as the starting point for specific discourse community identification. It suggests that language can play a modest but significant role in the acceptance of a RA by the community; however, most interestingly, it shows that authors need to balance the expectations of two audiences, both those in the immediate research community and those in the wider academic readership.

Appendix

See Tables 4, 5, 6, 7, 8, 9.

Table 4 "Endurance training" corpus: multi-dimensional analysis loadings

Dimension 1		Dimension 2		Dimension 3	
Non-numerical conclusion statement		Indicating results		Describing procedure	
Average word length	0.706	Verb—Past Participles	0.748	Present Participial WHIZ Deletion Relatives	0.569
Verbs—Base Form	0.711	Agentless Passives	0.586	Verbs—Gerund/Present Participle	0.502
Determiners	0.646	Private Verbs	0.548	Preposition/ Subordinating Conjunctions	0.411
Verb—Present Tense	0.605	Past Participial Clauses	0.467		
Possibility Modals	0.574	Adverbs	0.407		
Infinitives	0.529	Analytic Negations	0.386		
Nominalisations	0.460	Suasive Verbs	0.357		
Conjuncts	0.451				
Demonstratives	0.409				
Attributive Adjectives	0.407				
THAT verb complements	0.38				
Verbs—Past Tense	– 0.421				
Symbols	– 0.530				
Cardinal numbers	– 0.778				

Table 5 "Higgs boson" corpus: multi-dimensional analysis loadings

Dimension 1		Dimension 2		Dimension 3		Dimension 4		Dimension 5	
Mathematical terminology		Copular constructions		Passive constructions		Indicating tentative results		Subordination strategies	
Cardinal numbers	0.621	Verbs—present tense	0.725	Agentless passives	0.872	Verbs—base form	0.952	Verbs—gerund/present participle	0.754
Foreign words	0.542								
Symbols	0.420	BE as main verb	0.682	Verbs—past participles					
Type token ratio	0.380	Predicative adjectives							
Preposition/ subordinating conjunctions	0.352		0.620		0.761	Possibility modals	0.753	Present Participial WHIZ Deletion Relatives	0.635
Attributive adjectives	− 0.433	Other nouns	− 0.451						
Nominalisations	− 0.434								
Average word length	− 0.772								

Table 6 "Post-traumatic stress disorder" corpus: Multi-dimensional analysis loadings

Dimension 1		Dimension 2		Dimension 3	
Present relevance		Comparison of actions, events or states		Non-numerical description	
Verbs—present tense	0.851	BE as main verb	0.681	Average word length	0.836
Perfect aspect	0.588	Predicative adjectives	0.639	Attributive adjectives	0.524
Infinitives	0.397	Emphatics	0.428	Nominalisations	0.525
Public verbs	–			Other nouns	–
Verbs—past tense	0.383			Symbols	0.372
	–			Cardinal numbers	–
	0.719				0.420
					–
					0.590

Dimension 4		Dimension 5		Dimension 6	
Subordination strategies		Indicating tentative results		Passive constructions	
Present participial WHIZ deletion relatives	0.663	Verbs—base form	0.542	Verbs—past participle	0.810
	0.563	Type token ratio	0.541		
Suasive verbs	0.558	Possibility modals	0.536	Agentless passives	0.763
Verbs—gerund/present participle	0.379				
THAT verb complements					
		Phrasal coordination	– 0.393		

Table 7 "Endurance training" corpus: Spearman's correlation coefficients (r) of network statistics and linguistic features

	Tokens	AWL	TTR	AMP	ANDC	CAUS	CC	CD	CONC	COND	CONJ
Citations	0.34***	0	0.44***	0.04	− 0.14	− 0.13	− 0.01	0.06	0.05	− 0.07	− 0.06
Eigencentrality	0.12	− 0.13	0.24**	0.09	− 0.11	− 0.06	− 0.06	0.09	− 0.01	− 0.1	− 0.03
In-degree	0.13	− 0.23	0.11	− 0.04	− 0.06	− 0.12	0.07	0.15	0.01	− 0.18*	− 0.04
	DEMO	DEMP	DT	DWNT	EMPH	EX	FPP1	FW	GER	HDG	IN
Citations	0.03	0	− 0.18*	− 0.15	− 0.04	− 0.04	0.23	0.13	− 0.26**	− 0.07	0.38***
Eigencentrality	− 0.08	− 0.06	− 0.17*	0.04	− 0.04	− 0.02	− 0.02	0.12	0.05	− 0.07	0.15
In-degree	− 0.1	− 0.08	− 0.08	0.12	− 0.13	0.04	− 0.06	0.04	0.05	− 0.05	0.13
	INPR	JJ	LS	NEMD	NN	NOMZ	OSUB	PHC	PIN	PIT	PLACE
Citations	0.16	0.07	− 0.04	− 0.04	− 0.02	0	− 0.03	− 0.17*	− 0.03	− 0.07	− 0.06
Eigencentrality	0.06	0	0.02	− 0.04	0.04	− 0.11	0.05	− 0.19*	0.1	− 0.09	− 0.12
In-degree	− 0.03	− 0.16	0.07	− 0.05	0.1	− 0.12	0.07	− 0.08	0.07	− 0.18*	− 0.09
	POMD	POS	PRED	PRMD	PRP	QUAN	RB	RP	SYM	SYNE	THAC
Citations	0.07	− 0.22*	0.14	0.01	0.05	0.13	− 0.04	− 0.07	− 0.05	− 0.09	− 0.09
Eigencentrality	0.02	− 0.12	0.05	0	0.06	− 0.03	− 0.08	− 0.06	0.08	0.07	− 0.06
In-degree	− 0.04	− 0.17	0.05	− 0.13	− 0.05	0.02	− 0.13	− 0.1	0.06	0.11	− 0.1
	THVC	TIME	TO	TOBJ	TPP3	TSUB	VB	VBD	VBG	VBN	VPRT
Citations	0.01	0	− 0.13	− 0.06	0.01	0.47***	− 0.04	− 0.07	0.2	− 0.03	0.13
Eigencentrality	− 0.08	− 0.01	− 0.22**	0.03	− 0.05	0.3**	− 0.18*	0.06	0.06	− 0.08	− 0.07
In-degree	− 0.09	− 0.07	− 0.21	− 0.03	0.11	0.15	− 0.24	0.14	− 0.09	− 0.16	− 0.15
	WDT	WP	XX0	BEMA	BYPA	PASS	PASTP	PEAS	PIRE	PRESP	PRIV
Citations	− 0.06	0.11	0.08	0.1	0.11	− 0.29**	− 0.03	0.01	− 0.09	− 0.02	− 0.07
Eigencentrality	− 0.01	− 0.07	− 0.03	− 0.02	0.04	− 0.26**	− 0.01	− 0.11	− 0.13	0	− 0.09

Table 7 (continued)

	WDT	WP	XX0	BEMA	BYPA	PASS	PASTP	PEAS	PIRE	PRESP	PRIV
In-degree	− 0.07	− 0.08	− 0.03	− 0.07	− 0.05	− 0.16	− 0.06	− 0.18*	− 0.15	0.01	− 0.08
	PROD	PUBV	SERE	SMP	SPAU	SPIN	STPR	SUAV	THATD	WHCL	WHSUB
Citations	0.07	0.25**	0.02	− 0.15	− 0.08	0.18*	− 0.09	0.06	0.05	0.06	0.03
Eigencentrality	− 0.04	0.3**	0.17*	− 0.1	− 0.06	0.02	− 0.08	0.04	− 0.07	− 0.05	− 0.06
In-degree	0.05	0.14	0.13	− 0.18*	− 0.08	− 0.08	− 0.09	− 0.01	− 0.09	0.04	− 0.11
			WZPAST				WZPRES				
Citations			0.04				0.09				
Eigencentrality			0.07				0.07				
In-degree			0.06				− 0.1				

$n = 94$; Statistical significance: *** $p \leq 0.001$; ** $p \leq 0.05$; * $p \leq 0.1$

AWL average word length; *TTR* type token ratio; *AMP* amplifier; *ANDC* Independent clause coordination; *CAUS* causial adverbative subordinator; *CC* coordinating conjunction; *CD* cardinal number; *CONC* concessive adverbial subordinator; *COND* conditional adverbial subordinator; *CONJ* conjunct; *DEMO* demonstrative; *DEMP* demonstrative pronoun; *DPAR* discourse particle; *DT* determiner; *DWNT* downtoner; *EMPH* emphatic; *EX* existential *there*; *FPP1* first person pronoun; *FW* foreign word; *GER* gerund; *IN* preposition/subordinating conjunction; *INPR*indefinite pronoun; *JJ* attributive adjective; *LS* list item marker; *NEMD* necessity modal; *NN* other noun; *NOMZ* nominalisation; *OSUB* other adverbial subordinator; *PHC* phrasal coordination; *PIN* prepositional phrase; *PIT* pronoun *it*; *PLACE* place adverbial; *POMD* possibility modal; *POS* possessive ending; *PRED* predicative adjective; *PRMD* predictive modal; *PRP* personal pronoun; *QUAN* quantifier; *RB* adverb; *RP* particle; *SYM* symbol; *SYNE* synthetic negation; *THAC that* adjective complement; *THVC that* verb complement; *TIME* time adverbial; *TO* infinitive; *TOBJ that* relative clause on object position; *TPP3* third person pronoun; *TSUB that* relative clauses on subject position; *VB* verb, base form; *VBD* verb, past tense; *VBG* verb, gerund/present participle; *VBN* verb, past participle; *VPRT* present tense; *WDT wh−* determiner; *WP wh−* pronoun; *XX0* analytic negation; *BEMA be* as main verb; *BYPA by* passive; *PASS* agentless passive; *PASTP* past participial clause; *PEAS* perfect aspect; *PIRE* pied− piping relative clause; *PRESP* present participial clause; *PRIV* private verb; *PROD* pro-verb *do*; *PUBV* public verb; *SERE* sentence relative; *SMP seem/appear*; *SPAU* split auxiliary; *SPIN* split infinitive; *STPR* stranded preposition; *SUAV* suasive verb; *THATD* subordinator *that* deletion; *WHCL wh-* clause; *WHSUB wh-* relative clause on subject position; *WZPAST* past participial WHIZ deletion relatives; *WZPRES* present participial WHIZ deletion relatives

Table 8 "Higgs boson" corpus: Spearman's correlation coefficients (r) of network statistics and linguistic features

	Tokens	AWL	TTR	AMP	ANDC	CAUS	CC	CD	CONC	COND	CONJ
Citations	0.07	0.11	0.07	0.05	0.11**	− 0.06	0.06	− 0.08*	0.03	0.03	− 0.03
Eigencentrality	− 0.02	− 0.02	− 0.01	− 0.01	− 0.05	− 0.01	0	− 0.03	0.15***	0.05	− 0.04
In-degree	0.04	− 0.06	0.04	0.04	0.03	− 0.02	0.01	− 0.01	0.07	0.09**	− 0.05
	DEMO	DEMP	DPAR	DT	DWNT	EMPH	EX	FPP1	FW	GER	IN
Citations	− 0.01	− 0.02	0.18***	− 0.05	− 0.04	− 0.01	0.05	0.06	− 0.04	0.02	− 0.01
Eigencentrality	0.03	0.03	0	0.04	0.02	− 0.01	− 0.03	− 0.08*	− 0.05	− 0.02	− 0.02
In-degree	0	0.03	− 0.01	0.02	− 0.02	− 0.06	− 0.04	− 0.08*	− 0.02	− 0.03	0.03
	INPR	JJ	LS	NEMD	NN	NOMZ	OSUB	PHC	PIN	PIT	PLACE
Citations	0.08*	0.13**	− 0.03	0.02	− 0.15***	0.07	0.05	0.04	− 0.02	0.01	− 0.05
Eigencentrality	0.08*	0.05	− 0.04	0.02	− 0.02	− 0.05	− 0.04	− 0.01	0.07	0.02	0.02
In-degree	0.06	0.03	− 0.04	0.08*	− 0.02	− 0.08*	0.01	0	0.03	0	0.05
	POMD	POS	PRED	PRMD	PRP	QUAN	RB	RP	SYM	SYNE	THAC
Citations	− 0.03	− 0.01	0	0.03	0.01	0.03	0.06	0	− 0.02	− 0.01	− 0.01
Eigencentrality	0.09**	− 0.02	0.04	0.12**	0.08*	0.06	0.05	0.01	− 0.05	0.03	0.03
In-degree	0.07	0	0.06	0.13**	0.04	0.06	0.08*	0.06	0.01	0.01	0.01
	THVC	TIME	TO	TOBJ	TPP3	TSUB	VB	VBD	VBG	VBN	VPRT
Citations	0.01	− 0.01	0.04	− 0.03	0.03	0.02	0.05	0.11**	0.05	− 0.06	0.06
Eigencentrality	− 0.04	0.02	− 0.09**	− 0.01	0.12**	− 0.03	0	− 0.02	0.03	0.05	− 0.11**
In-degree	− 0.03	0.04	− 0.08*	0	0.08*	− 0.03	0.04	− 0.04	0.01	0.02	− 0.08*
	WDT	WP	XX0	BEMA	BYPA	PASS	PASTP	PEAS	PIRE	PRESP	PRIV
Citations	0.03	− 0.03	0.04	− 0.04	0.05	− 0.03	− 0.03	0.03	− 0.07	0.06	0.05
Eigencentrality	− 0.06	0.03	0.06	0.03	0.03	− 0.02	0.01	0.09**	− 0.06	0.12	− 0.03

Table 8 (continued)

	WDT	WP	XX0	BEMA	BYPA	PASS	PASTP	PEAS	PIRE	PRESP	PRIV
In-degree	− 0.05	− 0.04	0.04	0.03	0.02	− 0.01	0.03	0.08*	− 0.07	0.1	− 0.02
	PROD	PUBV	SERE	SMP	SPAU	SPIN	STPR	SUAV	THATD	WHCL	WHSUB
Citations	− 0.04	− 0.02	0.03	− 0.06	0.1**	0	0.1**	0.01	− 0.05	0.09**	0.06
Eigencentrality	− 0.02	0.08	− 0.03	− 0.08*	0.04	− 0.02	− 0.05	− 0.05	− 0.01	0.05	− 0.01
In-degree	− 0.04	0.06	− 0.02	− 0.07	0.08*	0	− 0.03	− 0.04	− 0.04	0.04	0

	WZPAST	WZPRES
Citations	− 0.03	0.02
Eigencentrality	0.03	− 0.01
In-degree	− 0.01	0.01

$n = 501$; Statistical significance: *** $p \leq 0.001$; ** $p \leq 0.05$; *$p \leq 0.1$

AWL average word length; *TTR* type token ratio; *AMP* amplifier; *ANDC* Independent clause coordination; *CAUS* causial adverbative subordinator; *CC* coordinating conjunction; CD cardinal number; *CONC* concessive adverbial subordinator; *COND* conditional adverbial subordinator; *CONJ* conjunct; *DEMO* demonstrative; *DEMP* demonstrative pronoun; *DPAR* discourse particle; *DT* determiner; *DWNT* downtoner; *EMPH* emphatic; *EX* existential *there*; *FPP1* first person pronoun; *FW* foreign word; *GER* gerund; *IN* preposition/subordinating conjunction; *INPR* indefinite pronoun; *JJ* attributive adjective; *LS* list item marker; NEMD necessity modal; *NN* other noun; *NOMZ* nominalisation; *OSUB* other adverbial subordinator; *PHC* phrasal coordination; *PIN* prepositional phrase; *PIT* pronoun *it*; *PLACE* place adverbial; *POMD* possibility modal; *POS* possessive ending; *PRED* predicative adjective; *PRMD* predictive modal; *PRP* personal pronoun; *QUAN* quantifier; *RB* adverb; *RP* particle; *SYM* symbol; *SYNE* synthetic negation; *THAC that* adjective complement; *THVC that* verb complement; *TIME* time adverbial; *TO* infinitive; *TOBJ that* relative clause on object position; *TPP3* third person pronoun; *TSUB that* relative clauses on subject position; *VB* verb, base form; *VBD* verb, past tense; *VBG* verb, gerund/present participle; *VBN* verb, past participle; *VPRT* present tense; *WDT wh–* determiner; *WP wh–* pronoun; *XX0* analytic negation; *BEMA be* as main verb; *BYPA by* passive; *PASS* agentless passive; *PASTP* past participial clause; *PEAS* perfect aspect; *PIRE* pied-piping relative clause; *PRESP* present participial clause; *PRIV* private verb; *PROD* pro-verb *do*; *PUBV* public verb; *SERE* sentence relative; *SMP seem/appear*; *SPAU* split auxiliary; *SPIN* split infinitive; *STPR* stranded preposition; *SUAV* suasive verb; *THATD* subordinator *that* deletion; *WHCL wh-* clause; *WHSUB wh-* relative clause on subject position; *WZPAST* past participial WHIZ deletion relatives; *WZPRES* present participial WHIZ deletion relatives

Table 9 "Post-traumatic stress disorder" corpus: Spearman's correlation coefficients (r) of network statistics and linguistic features

	Tokens	AWL	TTR	AMP	ANDC	CAUS	CC	CD	CONC	COND	CONJ
Citations	0.04	− 0.03	− 0.01	− 0.12	0.07	− 0.01	− 0.01	0.09	0.08	− 0.02	−0.09
Eigencentrality	− 0.15**	0.09	− 0.13*	0.11	0.05	− 0.03	0.04	− 0.03	− 0.07	− 0.03	− 0.07
In-degree	− 0.04	0.04	− 0.05	0.07	0.04	0.06	0.06	0.02	− 0.07	− 0.01	− 0.02
	DEMO	DEMP	DPAR	DT	DWNT	EMPH	EX	FPP1	FW	GER	IN
Citations	− 0.13*	0.08	0.21**	0.02	0.16**	0.02	− 0.05	− 0.04	− 0.09	− 0.11	− 0.09
Eigencentrality	0	0.02	0	0.07	− 0.06	− 0.03	− 0.06	− 0.04	0.06	0.12	− 0.12
In-degree	− 0.1	0.04	0.03	0.1	− 0.04	− 0.02	− 0.07	− 0.01	0.04	− 0.07	− 0.12
	INPR	JJ	LS	NEMD	NN	NOMZ	OSUB	PHC	PIN	PIT	PLACE
Citations	− 0.02	− 0.17**	0.27***	0.03	0.1	0.21**	0.04	0.1	0.06	0.04	0.08
Eigencentrality	− 0.02	− 0.07	− 0.02	0.02	0.11	− 0.04	− 0.01	− 0.15**	− 0.04	− 0.09	0.33***
In-degree	− 0.01	− 0.13*	0.11	0.06	0.19**	− 0.03	0.12	− 0.06	0	− 0.05	0.37***
	POMD	POS	PRED	PRMD	PRP	QUAN	RB	RP	SYM	SYNE	THAC
Citations	− 0.11	− 0.11	− 0.04	0.04	− 0.04	− 0.05	− 0.01	0.02	− 0.12	− 0.07	− 0.05
Eigencentrality	− 0.1	− 0.02	− 0.05	0.19**	− 0.03	− 0.04	− 0.09	− 0.02	0	− 0.05	− 0.03
In-degree	− 0.14*	− 0.07	− 0.1	− 0.04	− 0.06	− 0.02	− 0.1	0	− 0.07	− 0.06	− 0.06
	THVC	TIME	TO	TOBJ	TPP3	TSUB	VB	VBD	VBG	VBN	VPRT
Citations	− 0.11	− 0.09	0.11	− 0.05	0	− 0.04	0	− 0.1	− 0.06	− 0.07	0
Eigencentrality	0.04	− 0.1	− 0.09	− 0.03	− 0.08	0.09	− 0.12	0.05	0.01	0.04	0.08
In-degree	0.04	− 0.13	0	− 0.03	− 0.07	− 0.03	− 0.11	0.09	− 0.03	− 0.01	− 0.01
	WDT	WP	XX0	BEMA	BYPA	PASS	PASTP	PEAS	PIRE	PRESP	PRIV
Citations	− 0.08	− 0.07	− 0.02	− 0.11	− 0.02	− 0.13	− 0.04	0.08	− 0.03	− 0.05	− 0.04
Eigencentrality	0	0.03	0.16**	− 0.09	0.02	0.18	− 0.02	0.06	− 0.07	− 0.08	0.22

Table 9 (continued)

	WDT	WP	XX0	BEMA	BYPA	PASS	PASTP	PEAS	PIRE	PRESP	PRIV
In-degree	− 0.02	− 0.11	0.06	− 0.15	0.11	0.01	0.02	0.08	− 0.07	− 0.07	0.1
	PROD	PUBV	SERE	SMP	SPAU	SPIN	STPR	SUAV	THATD	WHCL	WHSUB
Citations	0.02	− 0.07	− 0.03	− 0.04	0.04	− 0.02	0.05	0	0.01	− 0.04	− 0.07
Eigencentrality	− 0.01	− 0.07	0.12	0.12	0.05	− 0.02	0	0.03	− 0.07	− 0.05	− 0.06
In-degree	− 0.01	− 0.06	0.07	0.14	0.08	0.01	0.01	0.04	− 0.08	− 0.04	− 0.09
				WZPAST				WZPRES			
Citations				0.03				0.02			
Eigencentrality				− 0.06				0.09			
In-degree				− 0.06				0.04			

$n = 176$; Statistical significance: *** $p \leq 0.001$; ** $p \leq 0.05$; *$p \leq 0.1$

AWL average word length; *TTR* type token ratio; *AMP* amplifier; *ANDC* Independent clause coordination; *CAUS* causial adverbative subordinator; *CC* coordinating conjunction; *CD* cardinal number; *CONC* concessive adverbial subordinator; *COND* conditional adverbial subordinator; *CONJ* conjunct; *DEMO* demonstrative; *DEMP* demonstrative pronoun; *DPAR* discourse particle; *DT* determiner; *DWNT* downtoner; *EMPH* emphatic; *EX* existential *there*; *FPP1* first person pronoun; *FW* foreign word; *GER* gerund; *IN* preposition/subordinating conjunction; *INPR* indefinite pronoun; *JJ* attributive adjective; *LS* list item marker; *NEMD* necessity modal; *NN* other noun; *NOMZ* nominalisation; *OSUB* other adverbial subordinator; *PHC* phrasal coordination; *PIN* prepositional phrase; *PIT* pronoun *it*; *PLACE* place adverbial; *POMD* possibility modal; *POS* possessive ending; *PRED* predicative adjective; *PRMD* predictive modal; *PRP* personal pronoun; *QUAN* quantifier; *RB* adverb; *RP* particle; *SYM* symbol; *SYNE* synthetic negation; *THAC that* adjective complement; *THVC that* verb complement; *TIME* time adverbial; *TO* infinitive; *TOBJ that* relative clause on object position; *TPP3* third person pronoun; *TSUB that* relative clauses on subject position; *VB* verb, base form; *VBD* verb, past tense; *VBG* verb, gerund/present participle; *VBN* verb, past participle; *VPRT* present tense; *WDT wh*-determiner; *WP wh*-pronoun; *XX0* analytic negation; *BEMA be* as main verb; *BYPA by* passive; *PASS* agentless passive; *PASTP* past participial clause; *PEAS* perfect aspect; *PIRE* pied-piping relative clause; *PRESP* present participial clause; *PRIV* private verb; *PROD* pro-verb *do*; *PUBV* public verb; *SERE* sentence relative; *SMP seem/appear*; *SPAU* split auxiliary; *SPIN* split infinitive; *STPR* stranded preposition; *SUAV* suasive verb; *THATD* subordinator *that* deletion; *WHCL wh-* clause; *WHSUB wh-* relative clause on subject position; *WZPAST* past participial WHIZ deletion relatives; *WZPRES* present participial WHIZ deletion relatives

Authors' contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Don Watson. The first draft of the manuscript was written by Don Watson and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Availability of data and material Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Agnoloni, T. (2014). Network Analysis of Italian Constitutional Case Law. In *Semantic Processing of Legal Texts (SPLeT–2014) Workshop Programme* (Vol. 91, No. F1, p. 24).

Barnett, G. A., Huh, C., Kim, Y., & Park, H. W. (2011). Citations among communication journals and other disciplines: A network analysis. *Scientometrics, 88*(2), 449–469. https://doi.org/10.1007/s11192-011-0381-2

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*. https://doi.org/10.13140/2.1.1341.1520

Begeny, J. C., & Greene, D. J. (2014). Can readability formulas be used to successfully gauge difficulty of reading materials? *Psychology in the Schools, 51*(2), 198–215. https://doi.org/10.1002/pits.21740

Behara, R. S., Babbar, S., & Smart, P. A. (2014). Leadership in OM research: A social network analysis of European researchers. *International Journal of Operations & Production Management, 34*(12), 1537–1563. https://doi.org/10.1108/IJOPM-08-2013-0390

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, D. (1992). The multi–dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities, 26*(5–6), 331–345. https://doi.org/10.1007/BF00136979

Busse, B., Gather, K., & Kleiber, I. (2016). Assessing the Connections between English Grammarians of the Nineteenth Century–A Corpus–Based Network Analysis. *Grammar and Corpora*, 435–442. https://doi.org/10.17885/heiup.361.509

Chen, L., Baird, A., & Straub, D. (2014). The evolving intellectual structure of the health informatics discipline: a multi–method investigation of a rapidly–growing scientific field. https://doi.org/10.2139/ssrn.2498225

Chen, B., Deng, D., Zhong, Z., & Zhang, C. (2020). Exploring linguistic characteristics of highly browsed and downloaded academic articles. *Scientometrics, 122*(3), 1769–1790. https://doi.org/10.1007/s11192-020-03361-4

Colladon, A. F., D'Angelo, C. A., & Gloor, P. A. (2020). Predicting the future success of scientific publications through social network and semantic analysis. *Scientometrics, 124*(1), 357–377. https://doi.org/10.1007/s11192-020-03479-5

Didegah, F., & Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. *Journal of Informetrics, 7*(4), 861–873. https://doi.org/10.1016/j.joi.2013.08.006

Dolnicar, S., & Chapple, A. (2015). The readability of articles in tourism journals. *Annals of Tourism Research, 52*, 161–166. https://doi.org/10.1016/j.annals.2015.03.003

Ebrahimi, S. F., & Chan, S. H. (2015). RA abstracts in applied linguistics and economics: Functional analysis of the grammatical subject. *Australian Journal of Linguistics, 35*(4), 381–397. https://doi.org/10.1080/07268602.2015.1070660

Epskamp, S., Cramer, A., Waldorp, L., Schmittmann, V., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software, 48*(4), 1–18. https://doi.org/10.18637/jss.v048.i04

Galvez, C. (2019). Thematic delimitation of the research in Linguistics and Communication through co–word analysis/Delimitacion tematica de la investigacion en Linguistica y Comunicacion mediante analisis de co-palabras. *Círculo De Lingüistica Aplicada a La Comunicación, 77*, 187–201. https://doi.org/10.5209/CLAC.63283

Gazni, A. (2011). Are the abstracts of high impact articles more readable? Investigating the evidence from top research institutions in the world. *Journal of Information Science*, *37*(3), 273–281. https://doi.org/10.1177/0165551511401658

Harrell Jr, F. E., & Harrell Jr, M. F. E. (2019). Package 'hmisc'. *CRAN2018*, *2019*, 235–236.

Hartley, J., Sotto, E., & Pennebaker, J. (2002). Style and substance in psychology are influential articles more readable than less influential ones?. *Social Studies of Science*, *32*(2), 321–334. https://doi.org/10.1177/0306312702032002005

Hewings, A., Lillis, T., Vladimirou, D. (2010). Who's citing whose writings? A corpus based study of citations as interpersonal resource in English medium national and English medium international journals. *Journal of English for Academic Purposes*, 9(2), June 2010, 102–115. https://doi.org/10.1016/j.jeap.2010.02.005

Huang, J. C. (2018). Marine engineering and sub–disciplinary variations: A rhetorical analysis of research article abstracts. *Text & Talk, 38*(3), 341–363. https://doi.org/10.1515/text-2018-0002

Humphries, M. D., & Gurney, K. (2008). Network 'small–world–ness': A quantitative method for determining canonical network equivalence. *PLoS ONE, 3*(4), e0002051. https://doi.org/10.1371/journal.pone.0002051

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly, 41*(2), 235–253. https://doi.org/10.1002/j.1545-7249.2007.tb00058.x

Jacques, T. S., & Sebire, N. J. (2010). The impact of article titles on citation hits: An analysis of general and specialist medical journals. *JRSM Short Reports, 1*(1), 2. https://doi.org/10.1258/shorts.2009.100020

Jamali, H. R., & Nikzad, M. (2011). Article title type and its relation with the number of downloads and citations. *Scientometrics*, 88(2), 653–661. https://doi.org/10.1007/s11192-011-0412-z

Jha, Y., & Welch, E. W. (2010). Relational mechanisms governing multifaceted collaborative behavior of academic scientists in six fields of science and engineering. *Research Policy, 39*(9), 1174–1184. https://doi.org/10.1016/j.respol.2010.06.003

Jiang, F. K., & Hyland, K. (2018). Nouns and academic interactions: A neglected feature of metadiscourse. *Applied Linguistics, 39*(4), 508–531. https://doi.org/10.1093/applin/amw023

Jin, B. (2018). A multidimensional analysis of RA discussion sections in the field of chemical engineering. *IEEE Transactions on Professional Communication, 61*(3), 242–256. https://doi.org/10.1109/TPC.2018.2817002

Kim, E. (2014). An analysis of move patterns in abstracts of social sciences RAs. *Journal of Korean Library and Information Science Society*, *45*(2), 283–309. https://doi.org/10.16981/kliss.45.2.201406.283

Knutas, A., Hajikhani, A., Salminen, J., Ikonen, J., & Porras, J. (2015). Cloud–based bibliometric analysis service for systematic mapping studies. In *Proceedings of the 16th International Conference on Computer Systems and Technologies* (pp. 184–191). ACM. https://doi.org/10.1145/2812428.2812442

Kortmann, B., & Szmrecsanyi, B. (2012). *Linguistic Complexity*. De Gruyter. https://doi.org/10.1515/9783110229226

Lei, L., & Yan, S. (2016). Readability and citations in information science: Evidence from abstracts and articles of four journals (2003–2012). *Scientometrics, 108*(3), 1155–1169. https://doi.org/10.1007/s11192-016-2036-9

Lu, C., Bu, Y., Dong, X., Wang, J., Ding, Y., Larivière, V., Sugimote, C., Paul, L., & Zhang, C. (2019). Analyzing linguistic complexity and scientific impact. *Journal of Informetrics, 13*(3), 817–829. https://doi.org/10.1016/j.joi.2019.07.004

McGrath, L., & Kuteeva, M. (2012). Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. *English for Specific Purposes, 31*(3), 161–173. https://doi.org/10.1016/j.esp.2011.11.002

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *Academy of Management Review, 23*(2), 242–266. https://doi.org/10.2307/259373

Nair, L. B., & Gibbert, M. (2016). What makes a 'good'title and (how) does it matter for citations? A review and general model of article title attributes in management science. *Scientometrics, 107*(3), 1331–1359. https://doi.org/10.1007/s11192-016-1937-y

Nini, A. (2019). The multi–dimensional analysis tagger. In Berber Sardinha, T. & Veirano Pinto M. (eds), *Multi–Dimensional Analysis: Research Methods and Current Issues,* 67–94, London; New York: Bloomsbury Academic. https://doi.org/10.5040/9781350023857.0012

Paiva, C. E., Lima, J. P. D. S. N., & Paiva, B. S. R. (2012). Articles with short titles describing the results are cited more often. *Clinics, 67*(5), 509–513. https://doi.org/10.6061/clinics/2012(05)17

Pho, P. D. (2008). RA abstracts in applied linguistics and educational technology: a study of linguistic realizations of rhetorical structure and authorial stance. *Discourse Studies*, *10*(2), 231–250. https://doi.org/10.1177/1461445607087010

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Revelle, W., & Revelle, M. W. (2015). Package 'psych.' *The Comprehensive R Archive Network, 337*, 338.

Spizzirri, L. (2011). Justification and application of eigenvector centrality. *Algebra in Geography: Eigenvectors of Network*. https://sites.math.washington.edu/~morrow/336_11/papers/leo.pdf

Subotic, S., & Mukherjee, B. (2014). Short and amusing: The relationship between title characteristics, downloads, and citations in psychology articles. *Journal of Information Science*, *40*(1), 115–124. https://doi.org/10.1177/0165551513511393

Swales, J. (1990). The concept of discourse community. *Genre analysis: English in academic and research settings*, 21–32.

Tseng, F. P. (2011). Analyses of move structure and verb tense of RA abstracts in applied linguistics. *International Journal of English Linguistics, 1*(2), 27. https://doi.org/10.5539/ijel.v1n2p27