Journal of Language and Literature
Vol. 22 No. 1, April 2022, pp. 238-254
DOI: 10.24071/joll.v22i1.4117
Available at https://e-journal.usd.ac.id/index.php/JOLL/index



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Thematic and Rhematic Progression in Scientific Writing: A Pilot Study

Alvin Ping Leong
alvin.leong@ntu.edu.sg
Language & Communication Centre, Nanyang Technological University, SINGAPORE

Abstract

The Hallidayan Theme-Rheme framework is commonly used to account for the message structure of language. Much has been done to investigate the message structure of different text types using the Hallidayan framework and Daneš's notion of thematic progression (TP). Similar studies targeting scientific research articles, however, are few, and the Rhemes in these studies are almost always sidelined. Diagrams capturing the development patterns of Themes and Rhemes at the whole-text level are also lacking. To address these gaps, this study compared the TP and rhematic progression (RP) patterns of 50 scientific research articles, adopting both a diagrammatic and quantitative approach. The quantification of TP and RP was based on the thematic-density index (TDI) and rhematic-density index (RDI), respectively. The results revealed that TDI was greater than RDI at all levels of the text. The TP and RP patterns were also different and distinct. Whereas the TP pattern comprised a simple-linear progression in the introduction section followed by a constant development in the rest of the article, the RP pattern was generally simple-linear in its shape. The observed TP and RP patterns capture not merely the message structure of scientific writing, but its communicative function. Further work involving more varied scientific texts is recommended to investigate whether these patterns are robust across disciplines.

Keywords: *rhematic progression; Rheme; scientific writing; thematic progression; Theme*

Article information

Received: 5 January 2022

Revised: 1 February 2022

Accepted: 14 February 2022

Introduction

The scientific research article plays a crucial role for scientists to share their work with the scientific community and beyond. It is the mainstay of scientific research. First appearing in a published form in 1665 in the *Journal des Sçavans* and the *Philosophical Transactions of the Royal Society of London* (Larivière et al., 2015), its form and style of writing have evolved over the years in lockstep with the increasing emphasis on scientific rationality (Bazerman, 1988; Dimković-Telebaković, 2012; Ding, 1998; Gross et al.,

2002). Today, few have problems recognizing such articles as a specialized type of writing, often basing their judgments on the presence of technical terms, equations, and tables/graphs in them (Alley, 2018). Beyond these obvious markers, however, it is less easy to articulate broader structural features that characterize research articles in general. Cargill and O'Connor (2009), for instance, have observed variations in the rhetorical structure of scientific articles, noting that whereas shorter scientific articles follow the more classical introduction-methodology-results/discussion order, molecular biology articles tend to highlight the results and discussion, placing them before the methodology section. Journals may also have their own preferred article sections for the sake of readability and uniformity.

Such structural variations are perhaps only to be expected, given the diverse disciplines and research methodologies in the sciences. It is important to note, though, that in scientific writing, like any other form of writing, language remains a tool for communication. Scientific texts are thus no different from other text types in this one fundamental aspect—that language pushes the discourse forward. How this is done in scientific articles can be investigated from at least two perspectives. The first concerns the rhetorical segments in the major sections of the text. Much work has been done in this respect in various disciplines, ranging from the pioneering effort of Swales (1981) on the introduction section to the studies by Darabad (2016) on the structure of abstracts in linguistics, mathematics, and chemistry; Williams (1999) on the results section of medical research articles; Kanoksilapatham (2005) on the rhetorical moves in biochemistry research articles; and Cronin et al. (1992), and Costas and van Leeuwen (2012) on the acknowledgement segment in scholarly writing.

These studies have provided valuable insights into the organization of scientific articles; the limitation, however, is that they regard the major sections as a given. For instance, while Kanoksilapatham (2005) identified 15 distinct moves in his corpus at the whole-text level, these are nevertheless grouped by sections—"three moves for the Introduction section, four for the Methods section, four for the Results section, and four for the Discussion section" (p. 269). As we have seen in the preceding paragraph, scientific articles exhibit variations in the presence and ordering of these sections, not all of which are entirely of the writers' choosing. This naturally raises the question of whether there exists a structural norm in scientific writing that goes beyond such section boundaries.

A possible answer to this question is offered by the second perspective. Here, we track the development of the *message* components of each clause through the text. These message components, as conceptualized in the Hallidayan framework (Halliday & Matthiessen, 2014), are Theme and Rheme (these terms are capitalized, following the Hallidayan convention). The former serves as the point of departure, and the latter is the remainder of the clause. This second perspective, that is to say, looks at the message, rather than the rhetorical, structure of the text. It relies on the early work of Daneš (1970, 1974) on thematic progression (TP), which tracks how each message component links to other message components in the text. TP thus shows how the Theme and Rheme of each clause develop the message in the larger text.

Past studies investigating the TP patterns in scientific articles, however, are not many. Early attempts include the work of Dubois (1987) on biomedical articles, and Nwogu and Bloor (1991) and Williams (2009) on medical articles. These studies, though, neglected to address the *global* TP pattern of articles. The work of Williams (2009), for instance, is restricted to only the discussion section. Some effort was made to correct this in the work of Leong (2015) and Leong et al. (2018), both of which highlighted an "anchored" development of Theme at the whole-text level (the corpus in each study comprised biology-related articles). While these recent studies are a promising step forward, they (and related studies in general) suffer from a further limitation—the tendency to focus on Theme, rather than Theme *and* Rheme. Indeed, neither Leong (2015) nor Leong et al. (2018) considered Rheme at all in their analyses. Focusing on only Theme therefore leaves us with an incomplete picture

of how scientific writing pushes the discourse forward.

To address these research gaps, this study investigated the progression patterns of both Theme and Rheme in a corpus of 50 scientific research articles in the field of nanotechnology. It used a Microsoft Excel-based semi-automated template to generate Theme-Rheme diagrams to capture the global patterns of the articles. To the best of my knowledge, there has not been any prior study done involving rhematic progression. In this light, the work reported here is a pilot study. It is hoped, nevertheless, that it will add to our understanding of how the scientific message in research articles is typically organized.

Theme, Rheme, and Related Studies

The Message Structure of the Clause

Halliday's view on the message structure of the clause was first articulated in a series of seminal papers published in the 1960s (Halliday, 1967, 1968). It forms part of Halliday's functional theory of language (Halliday & Matthiessen, 2014), which regards language as performing three essential functions, what he refers to as "metafunctions". These metafunctions are the *ideational* (which construes one's experience of the world, real or imagined), the *interpersonal* (which establishes interpersonal relations between or among discourse participants), and the *textual* (which packages these experiences and interpersonal relations into a coherent text). The message structure of language belongs to the textual metafunction.

According to the framework, each clause has a two-part message structure, comprising a Theme and a Rheme, in that order. Three types of Themes—*textual*, *interpersonal*, and *topical*—are distinguished, reflecting the three metafunctions recognized in the larger framework. *Textual Themes* serve a connecting function and are typically realized by conjunctions and clause-initial conjunctive adjuncts. *Interpersonal Themes* reflect not only the encoder's attitudes but also the nature of language as a means of interaction. They typically comprise modal adjuncts and the finite operators of verb phrases. *Topical Themes* are the most important of the three Theme types; unlike textual and interpersonal Themes, which are optional, topical Themes are obligatory in all finite clauses. They are realized by the first participant, first main verb, or first circumstantial element in the clause. Such elements serve a crucial function in grounding the clausal message, serving as its point of departure. Halliday and Matthiessen (2014) argue that without a topical Theme, "the clause lacks an anchorage in the realm of experience" (pp. 111–112). The topical Theme ends the thematic portion of the clause; the remainder of the clause, which develops the topical Theme, is the Rheme. The linguistic elements realizing each Theme type are summarized in Table 1.

The message structure, as conceived in the Hallidayan framework, is based on the grammatical clause as the basic unit of analysis. One way to extend this framework beyond the clause is offered by the notion of thematic progression, to which we now turn.

Thematic Progression

Thematic progression (TP) was proposed by Daneš (1970, 1974) to track how the Theme and Rheme of each clause are semantically related to those of other clauses in the text. Tracing the development of Themes and Rhemes this way reveals what Daneš (1974) terms "the skeleton of the plot" (p. 114). It offers us a way to see how the message moves from clause to clause, thereby illustrating the way the discourse is pushed forward. Daneš's investigation of Czech scientific and professional texts led him to propose several canonical TP patterns, two of which are the *simple-linear* TP and the *constant* TP. These are illustrated in Figures 1–2, respectively, using examples taken from the corpus. In all the examples used in this paper, independent clauses are separated using double vertical lines ||, and topical Themes are highlighted in boldface (textual and interpersonal Themes are omitted as they are optional). Ellipsed Themes are enclosed within square brackets []. In the diagrams, Themes are represented by 'T' and Rhemes by 'R'; the arrows indicate how the Themes and Rhemes are linked in terms of content. In Figure (1), for example, '$R_1 \rightarrow T_2$' means that the Rheme of the first clause (R_1 = "composite solid electrolyte") and the Theme

of the second clause (T_2 = "composite SSE") share the same referent.

(1) || Here, **we** propose the design of an ultrathin, high-performance polymer–polymer composite solid electrolyte for all-solid-state Li batteries. (Fig. 1a) || **The composite SSE** [solid-state electrolyte] is made of a robust, nonflammable host with vertically aligned nanochannels and Li-ion conductive SPE fillers. || **The high modulus host** prevents potential dendrite penetration || […]

$T_1 \rightarrow R_1$
$\downarrow$
$T_2 \rightarrow R_2$
$\downarrow$
$T_3 \rightarrow R_3$

Figure 1. Simple-linear TP, based on example (1)

(2) || **Secondary nanoplastics** are also present in the environment, || and **[secondary nanoplastics]** range from tyre wear to fragmented mismanaged waste. || **These sources of plastic** will make their way to the WWTP through road runoff in locations that have combined sewer systems. ||

$T_4 \rightarrow R_4$
$\downarrow$
$T_5 \rightarrow R_5$
$\downarrow$
$T_6 \rightarrow R_6$

Figure 2. Constant TP, based on example (2)

Table 1. Linguistic elements realizing textual, interpersonal, and topical Themes (adapted from Halliday and Matthiessen, 2014, pp. 105–114)

Theme	Linguistics elements
Textual Theme	Continuatives Conjunctions or conjunctive adjuncts *Wh*- Relatives
Interpersonal Theme	Vocatives Modal adjuncts Finite operators *Wh*- question words/phrases (content interrogatives)
Topical Theme	First participant, first circumstantial adjunct, or first main verb

TP studies involving scientific articles are not common. Dubois's (1987) study on biomedical texts is an early effort, and her work led to the identification of two other TP patterns—*multiple* and *gapped* developments, the former referring to a particular Theme being developed in multiple patterns, and the latter, to a pattern being interrupted by a short passage. The work of Nwogu and Bloor (1991) is another valuable contribution. Their corpus involved, in part, medical research articles, and the study revealed that the constant TP was commonly found in them.

Two issues are apparent from a quick survey of these early TP studies and the TP diagrams in Figures 1–3. The first, as alluded to in the preceding section, is the focus on Theme. It is the arrangement of Themes, not Rhemes, that gives each TP pattern its name. Such a practice, however, runs the risk of backgrounding rhematic patterns, which may well be as insightful as thematic patterns in

revealing the global message structure of the text. In the literature, the earliest attempt to draw attention to rhematic patterns can be found in the work of Enkvist (1974), who proposed two other concatenations—*Rheme iteration* and *Rheme regression* (Figure 3). However, there has been little scholarly interest on rhematic patterns since then.

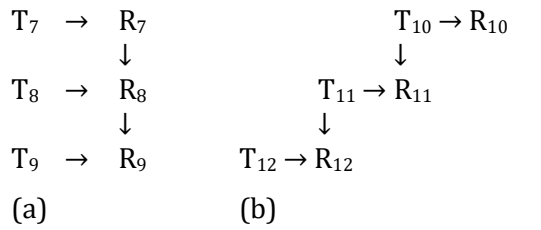


Figure 3. (a) Rheme iteration, and (b) Rheme regression

In fact, the very notion of Rheme itself is rarely addressed in depth. Available studies tend to focus on other languages, such as Russian (Khaldoyanidi & Morel, 2013), Swedish (Udilova, 2009), and ancient Greek (Viti, 2008), or on grammatical issues (Liubov, 2011).

The second issue concerns the representation of TP patterns. Figures 1–3 suggest that TP is essentially a *diagrammatic* representation of the text. Producing TP diagrams at the whole-text level, unfortunately, is complicated and effortful. It is perhaps for this reason that related studies have tended to include only TP diagrams for short passages (e.g., McCabe, 1999). The preference, instead, is to compute and compare the totals of the identified TP patterns in the text (e.g., Williams, 2009). While such a quantitative approach is convenient, it hides the macro, "skeletal" shape of the text, something which a diagrammatic representation is better suited to capture. It is also constrained by the fact that TP patterns must be established *beforehand* (in order for them to be identified and counted); Themes and Rhemes that do not fit these patterns at the local level may thus be left out of the computation. This, however, can be rather misleading because such elements may nevertheless contribute to the global pattern in some way.

Recent Developments

As noted from the second issue above, producing TP diagrams at the whole-text level presents practical challenges for the analyst. In response to this, Leong (2015) and Leong et al. (2018) used Microsoft Excel to help them generate simplified text-level TP diagrams of the biology-related research articles in their corpora. Confining their work to only topical Themes, the term TP tracked the development of *only* Themes in the text, thus differing from Daneš's original notion (see Figures 1–2). The authors discovered a general linear progression of Themes in the introduction section, followed by a constant development in the rest of the article. They referred to the constant TP as an "anchored" development, underscoring Halliday's (1970) description of Theme as "the peg on which the message is hung" (p. 161). This broad structure is exemplified in Figure 4, where the Themes of each clause are indicated by the black dots. The X-axis relates to the semantic content of each Theme (in sequential order of occurrence in the text), and the Y-axis represents the individual clauses in the text, beginning with the first clause at the top of the diagram. Further details about the construction of such diagrams are given in 'Method of Analysis' in the 'Methodology' section.

As Figure 4 illustrates, the TP pattern is a *general* development as seen from the macro level. Although there are frequent thematic interruptions and outlying Themes at the local level, the clear shape that emerges is an initial simple-linear development followed by a constant development.

The findings of Leong (2015) and Leong et al. (2018), which are preliminary and thus tentative in nature, certainly need to be further verified and compared using research articles from other scientific disciplines to obtain a fuller understanding of scientific communication in general. The authors' focus on only Theme in their work also leaves the issue concerning Rheme unresolved.

Journal of Language and Literature
ISSN: 1410-5691 (print); 2580-5878 (online

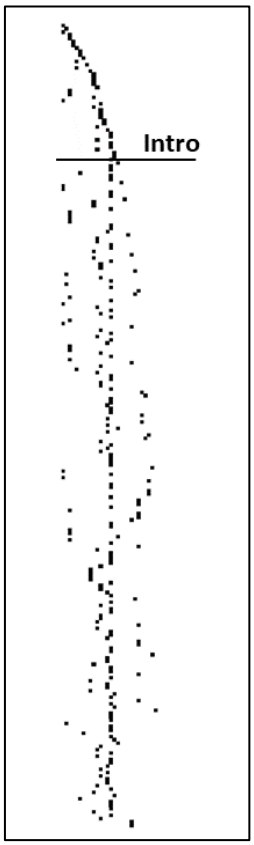


Figure 4. Text-level message structure, illustrating a general simple-linear TP followed by a general constant TP (adapted from Leong et al., 2018, p. 300)

To address these gaps, this present work expanded on these recent studies to investigate the message structure of a corpus of scientific research articles in the field of nanotechnology. It used the narrower definition of TP to refer to only the semantic development of topical Themes in the text, and introduced *rhematic progression* (RP) to refer to the development of Rhemes. TP and RP diagrams at the whole-text level were generated using an improved Microsoft Excel tool, and the diagrams were quantified for statistical testing using a measure first proposed in Leong (2016) and later tested in Leong et al. (2018). These and other methodological details are presented in the next section.

Methodology

Corpus

The corpus comprised 50 research articles from the journal *Nature Nanotechnology*. All the articles were published in 2019, and were the most recent articles at the time of analysis. The journal and articles were selected for two reasons. First, the journal is highly esteemed. According to *Scimago Journal and Country Rank* (https://www.scimagojr.com/) for the year of assessment 2019, the journal was ranked first in the field of "engineering", second in "chemical engineering" and "physics and astronomy", and third in "materials science". Second, as indicated in its ranking categories, the journal publishes articles from a variety of disciplines, both scientific and technical, in the broad field of nanotechnology; this is helpful in offering us a glimpse of the message structure of scholarly writing in not just one or two topic areas, but diverse areas, ranging from semiconductors to pesticides and drug-delivery systems.

The articles had a total of 162,080 (M = 3,241.60 words) words and 6,680 (M = 137.20) independent clauses. The basic unit of analysis was the independent clause; dependent clauses, including embedded clauses, were not analyzed for Theme and Rheme. As Fries and Francis (1992, p. 47) note, focusing on only the independent clause allows the analyst to "discern the method of development and thematic progression of a text" more easily, since "the structure of beta [dependent] clauses, including their thematic structure, tends to be constrained by the alpha [independent] clauses". This is also the common practice adopted in other text-based studies (e.g., McCabe, 1999; Williams, 2009).

Method of Analysis

Each article was first divided into independent clauses, and the topical Theme and Rheme of each clause were identified. Semantic labels were assigned to reflect the semantic content of each message component, and new labels were added as necessary. These semantic labels were tracked using a Microsoft Excel template designed by Leong (2019). Each row in the template represented one independent clause (and any

dependent/embedded clause(s) attached to it), and each column represented, in sequence, the semantic labels assigned to the Themes and Rhemes.

In the analysis, determining the semantic content of each topical Theme was a fairly straightforward matter since each topical Theme is realized by one grammatical constituent (i.e., the first participant, main verb, or circumstantial element). Determining the semantic content of the Rheme, on the other hand, was more complicated because the Rheme, by definition, is the remainder of the clause following the topical Theme. As this remaining segment may comprise a number of constituents, a decision had to be made to select a *core* constituent to represent the semantic load of the Rheme. It was decided that the participant in the Rheme would serve as this core constituent. Halliday and Matthiessen (2014, p. 154), in fact, regard clausal participants as being "inherent", as opposed to circumstantial adjuncts, which are "attendant". More specifically,

> Our most powerful impression of experience is that it consists of a flow of events, or 'goings-on'. This flow of events […] is modelled as a figure—a figure of happening, doing, sensing, saying, being or having […] All figures consist of a process unfolding through time and of participants being ***directly*** involved in this process in some way; and in addition there may be circumstances of time, space, cause, manner or one of afew other types. These circumstances are not directly involved in the process; rather they are attendant on it. (Halliday & Matthiessen, 2014, p. 214; my emphasis)

Hence, where the Rheme was concerned, the default principle was to select the participant when determining the semantic content. In the absence of any rhematic participant, the circumstantial adjunct or the main verb, in that order, was then selected.

An example of the analysis, using the first seven clauses of a sample article from the corpus, is presented in (3) and the accompanying Figure 5. In (3), the numbers enclosed in square brackets are reference numbers, and the key words of the core rhematic elements are underlined. The black and grey squares in Figure 5 represent Themes and Rhemes, respectively, and "TR" is used to indicate that the Theme and Rheme in the same clause are equated, as in (3[003]). Figures 5(b–c), representing the TP and RP of (3), are derived from Figure 5(a).

(3)

[001] **The physical confinement of water at the nanoscale** can play a major role in controlling its properties, with fundamental implications in physical, chemical, geological and biological phenomena.

[002] Not surprisingly, **the mobility of nanoconfined water along with its behaviour at interfaces** has attracted widespread attention.

[003] In this regard, **the nature of the interface and the geometric details of the confining surface** are key parameters.

[004] In particular, **confinement in the nanometre range** can inhibit the arrangement of water molecules into an ice structure,

[005] and **[confinement in the nanometer range]** thereby prevent crystallization at subzero temperature

[006] and **[confinement in the nanometer range]** create a state of amorphous water.

[007] **Confinement within soft interfaces**, such as those formed by the self-assembly of surfactants in an aqueous environment, was suggested as a model for confined water in a cellular environment.

As shown in Figures 5(b–c), the TP and RP patterns are markedly different; the pattern in

Journal of Language and Literature
ISSN: 1410-5691 (print); 2580-5878 (online

the former takes a general constant shape whereas the latter is clearly simple-linear. For a more objective comparison of diagrams, Leong (2016) proposed a thematic-density index (TDI), which is the quotient of the number of clauses by the number of semantic labels corresponding to the Themes in the text. The formula is given in (4):

$$(4)\ \mathrm{TDI}_i = \frac{(\text{Number of clauses})_i}{(\text{Number of semantic labels})_i}$$

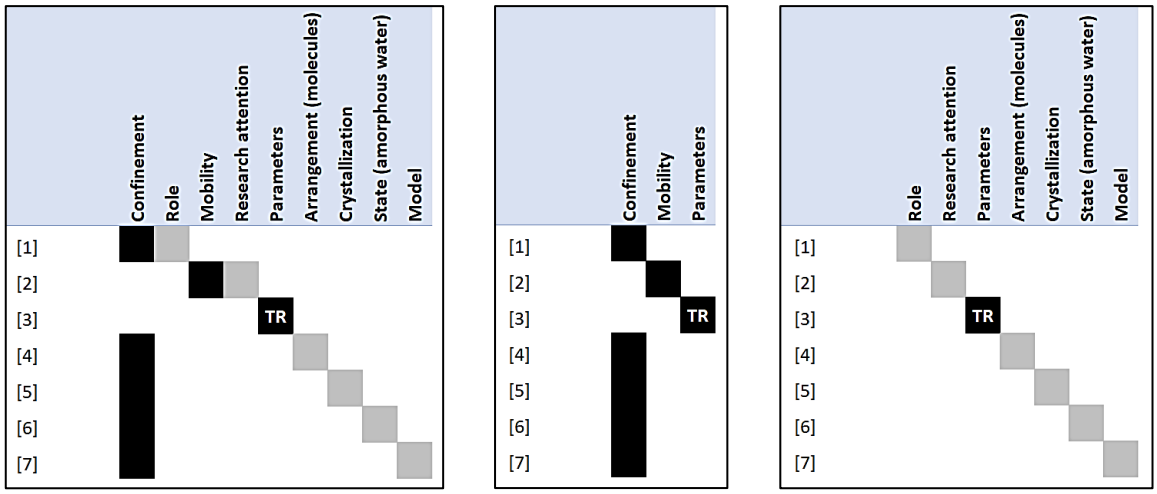


Figure 5. Example analysis of (3), displaying the (a) combined progression of Themes (black) and Rhemes (grey), (b) TP only, and (c) RP only; the label "TR" refers to an instance where Theme is equated with Rheme

In terms of Figure 5(b), this is essentially a matter of dividing the number of rows by the number of columns. The possible values of TDI therefore range from 1 to the total number of clauses in the text. These polar values reflect a constant TP and a simple-linear TP, respectively. A text with a higher TDI implies that it is thematically "dense", i.e., the topical Themes tend to cluster around few semantic labels. By contrast, a text with a lower TDI has Themes that are dispersed across more semantic labels. As Leong (2016) excluded Rhemes in his study, the TDI is confined to only Themes. However, the same formula in (4) can be easily extended to Rhemes as well. This therefore gives us both a TDI and an RDI (rhematic-density index) for any one text. In the case of Figures 5(b–c), the TDI is $\frac{7}{3} = 2.33$, and the RDI is $\frac{7}{7} = 1$. These numbers are useful in allowing for differences between TDIs and RDIs to be statistically tested.

Statistical Analysis

The *Real Statistics Resource Pack* for Microsoft Excel (Zaiontz, 2020) was used for all statistical tests (Student's *t*-test, two-tailed). The significance level for all tests was α=0.05.

Results and Discussion

Broad Findings

The broad statistics for the macro structure of the research articles in the corpus are given in Table 2. The TDI and RDI were computed for the article as a whole ($\mathrm{TDI}_{\mathrm{whole}}$, $\mathrm{RDI}_{\mathrm{whole}}$), and for two major segments of each article—the introduction section ($\mathrm{TDI}_{\mathrm{intro}}$, $\mathrm{RDI}_{\mathrm{intro}}$) and the rest of the article ($\mathrm{TDI}_{\mathrm{rest}}$, $\mathrm{RDI}_{\mathrm{rest}}$). No fair comparisons were possible involving the other sections since the rhetorical structures of the articles were not the same, given the different sub-disciplines included in the journal.

Table 2. TDI and RDI at the whole-text level and in two major segments

Whole Text		
TDI_{whole}	3.38 (CV = 0.23)	
RDI_{whole}	2.71 (CV = 0.26)	$t(98) = 4.52$, $p = 1.76e^{-5}$
Introduction Section		
TDI_{intro}	1.82 (CV = 0.19)	
RDI_{intro}	1.42 (CV = 0.16)	$t(98) = 6.71$, $p = 1.25e^{-9}$
Rest of Article		
TDI_{rest}	3.19 (CV = 0.23)	
RDI_{rest}	2.57 (CV = 0.26)	$t(98) = 4.34$, $p = 3.43e^{-20}$
Introduction Section vs. Rest of Article		
TDI_{intro}	1.82 (CV = 0.19)	
TDI_{rest}	3.19 (CV = 0.23)	$t(98) = 11.71$, $p = 2.45e^{-20}$
RDI_{intro}	1.42 (CV = 0.16)	
RDI_{rest}	2.57 (CV = 0.26)	$t(98) = 11.34$, $p = 1.57e^{-19}$

Table 2 also reports the coefficient of variation (CV) for TDI and RDI at the whole-text level, the introduction section, and the rest of the article. The CV, which is the ratio of the standard deviation to the mean of each distribution, measures the dispersion of data points around the mean. As can be seen in Table 2, the CV values range from 0.16 (RDI_{intro}) to 0.26 (RDI_{whole}, RDI_{rest}), suggesting low variability.

All the observed differences between TDI and RDI were highly significant. This highlights not only a difference between the developments of Themes and Rhemes at the whole-text level, but also at the introduction section vis-à-vis the rest of the article. The differences between the TDIs and RDIs, in fact, were also generally uniform—the TDIs were roughly 1.26 as large as the RDIs in the various scenarios.

The broad results therefore indicate that the topical Themes in the corpus were clustered around fewer semantic labels than the Rhemes. The higher values of TDI_{whole} and TDI_{intro} suggest greater thematic density, implying a TP pattern that is distinctly different from an RP pattern. The TP and RP diagrams for three illustrative texts are given in Figure 6. For each research article, the TP and RP diagrams are positioned next to each other for easier comparison. The diagrams follow the conventions as described for Figures 4 and 5. These examples are representative of the larger corpus; as indicated by the CV values in Table 2, there is low variability across the texts in the corpus.

Journal of Language and Literature
ISSN: 1410-5691 (print); 2580-5878 (online

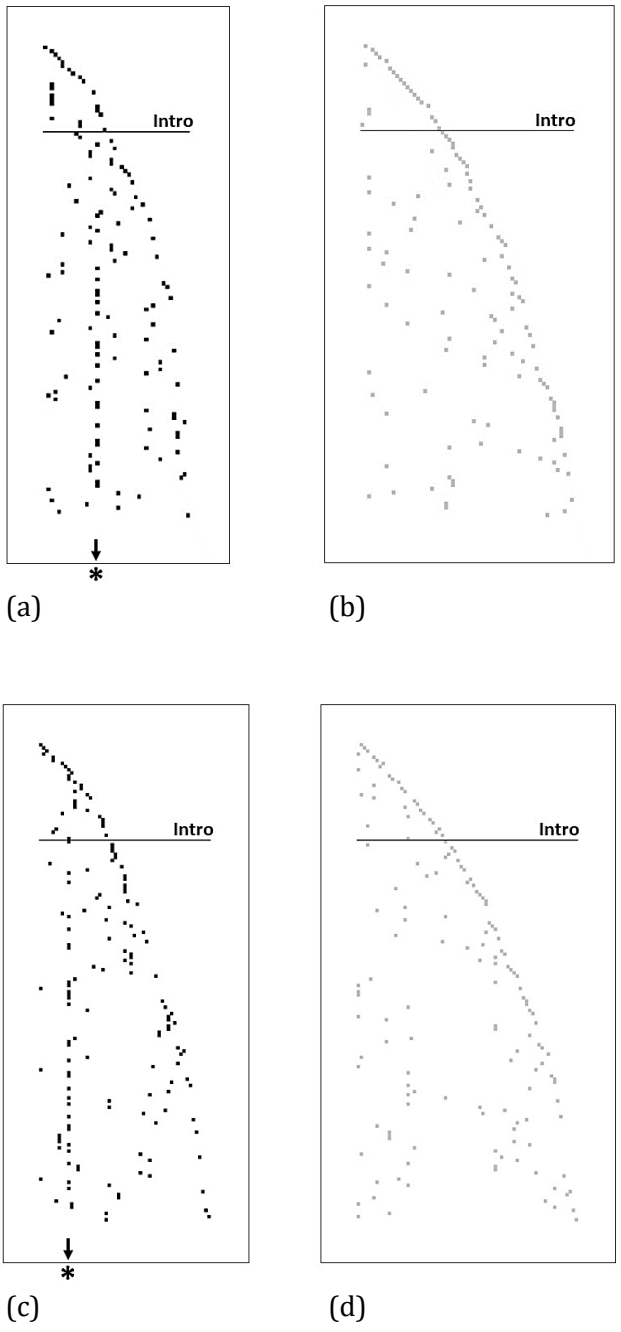
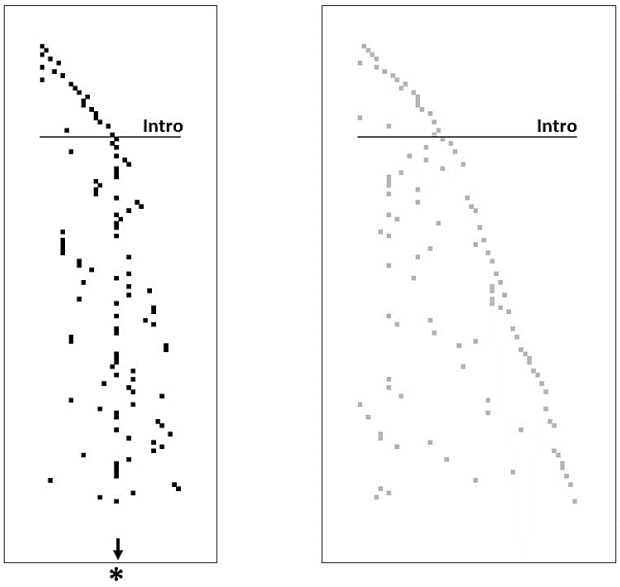


Figure 6. Example TP and RP diagrams from three representative articles



(e) (f)

Figure 6 (cont'd). Example TP and RP diagrams from three representative articles

The TP diagram takes the form of two distinct shapes—a simple-linear shape in the introduction section followed by a constant development in the rest of the article. The vertical line representing the constant development is marked out with an asterisk in Figure 6(a, c, e). By comparison, the pattern involving Rhemes is more progressive. I discuss more fully these thematic and rhematic shapes, and what each implies, in the following two sections.

Themes and Thematic Progression

The general shape of the TP diagrams in Figure 6 concurs well with past observations about both the introduction section and the scientific article as a whole. In the introduction section, the simple-linear progression of Themes represents a gradual narrowing of ideas toward the focus, mirroring the general arrangement of information as presented in Swales's (1981, 1990) model. They resemble what Swales and Feak (2004, p. 44) note about "general-specific texts", which "move from broad statements to narrower ones". An example of such a message flow is seen in (5) below, taken from the first seven independent clauses of an article on how nanomaterials can result in gaps in the endothelial walls of blood vessels.

(5)

[001] **Cancer metastasis** is a phenomenon in which cancer cells disseminate from a primary tumour to eventually grow at distant sites.

[002] **The metastatic stage of many solid tumour cancers** usually presents a poor prognosis and likewise accounts for the vast majority of cancer-related mortality (~90%).

[003] **Central to the pathophysiology of metastasis** is the intravasation and extravasation of cancer cells through disrupted blood vessels.

[004] **This** highlights the importance of intact vasculature against isolated but migratory cancer cells.

[005] **Cancer nanotechnology** offers numerous possibilities in diagnosing and treating cancers due to their many possible and interesting interactions.

[006] Since some nanoparticles (NPs) could induce endothelial leakiness (NanoEL), **cancer nanomedicines**, designed to kill

the tumour, may also unintentionally induce leakiness of the tumour vasculature, thereby lowering the barrier for intravasational entry of surviving cancer cells into the circulation.

[007] **NanoEL** came to light when certain nanomaterials disrupted endothelial cell–cell interactions by binding to critical adherens junction proteins such as vascular endothelial-cadherin (VE-cadherin).

As can be seen, the Themes in the independent clauses narrow from cancer metastasis to the treatment of cancers using nanotechnology to the unfortunate consequence of nanomaterials causing endothelial leakiness, the focus of the authors' work.

Beyond the introduction section, the constant thematic development also agrees well with the findings of Leong et al. (2018) and Nwogu and Bloor (1991), who noticed the same trend in biology and medical research articles, respectively. This provides further suggestive evidence of a thematic pattern that appears to be common in scientific writing. The constant development of topical Themes indicates the use of a central idea (or a small set of ideas) as the point of departure of the textual message. The TP diagrams revealed that these points of departure—marked out with asterisks in Figure 6(a, c, e)—tended to be first-person pronouns referencing the authors or their own work. This is further exemplified in (6), taken from the same article used in (5) above.

(6)

[070] **One alternative explanation** is that NPs could increase the intrinsic migratory ability of breast cancer cells directly without the involvement of NanoEL.

[071] **We** checked that possibility with various migration assays using cancer cells that were exposed to TiO_2, SiO_2 and Au NPs.

[072] **We** found no obvious changes in migration and epithelial–mesenchymal transition (EMT) markers even after 24 h of treatment (Supplementary Figs. 18 and 19).

[073] Combined with Fig. 4a, **we** can conclude that increasing the dose of TiO_2 NPs may have increased the NanoEL effect and likewise increased intravasation of the tumour cells without changing MDA-MB-231 cellular behaviour.

Here, the first-person pronouns do not simply indicate what the authors did (6[071–072]), but also what they inferred from their own work (6[073]). This is entirely consistent with the findings of Leong et al. (2018, p. 306), who noted that "[t]hrough the pronoun "we," the authors claim responsibility for both methodological decisions and the expression of opinions or arguments in the rest of the article". Martínez (2005, p. 182) adds that this tendency to articulate in the first person could represent a trend in scientific writing toward "authorial intervention, argumentation, and personalization". As the basic function of the Theme is to establish the point of departure, or the "ground from which the clause takes off" (Halliday, 1994, p. 38), this finding also highlights the foundational, important role authors play in scientific discourse. At its core, scientific writing is less about the topic of investigation, and more about the authors' research efforts and contributions in relation to that topic. Indeed, nothing can be researched without the researchers, and it is perhaps fitting that the message structure of scientific writing captures this basic truth.

The use of the first-person pronouns, interestingly, also reflects the changing language norms in scientific writing. Whereas the passive voice was common in scientific writing in the past, studies by Leong (2014) and Banks (2017) have shown that modern scientific writing prefers an active-voice style to make the writing more accessible to readers. For instance, top journals, such as *Nature* (including *Nature Nanotechnology*), have specific guidelines regarding the use of the grammatical voice:

> Nature journals prefer authors to write in the active voice ("we performed the experiment…") as experience has shown that readers find concepts and results to be conveyed more clearly if written directly. (*Nature*, 2020)

At the macro level, then, we see from the TP diagrams a reflection of two important functions of Themes in scientific writing—the Themes (a) guide the reader toward the research focus in the introduction section, and (b) position the researchers as the initiators and agents of the research work in the rest of the article.

Rhemes and Rhematic Progression

Halliday and Matthiessen (2014, p. 89) characterize the Rheme as "the part in which the Theme is developed". As the development of Themes can proceed in a number of ways, the corresponding RP pattern, technically, can also take a variety of forms. Where scientific writing is concerned, however, the present analysis revealed a consistent simple-linear RP shape (see Figure 6(b, d, f)). Although RDI_{intro} (M = 1.42) is lower than RDI_{rest} (M = 2.57; $t(98) = 11.34$, $p = 1.57e^{-19}$), it should also be borne in mind that RDI_{rest} is lower than TDI_{rest} (M = 3.19; $t(98) = 4.34$, $p = 3.43e^{-20}$). Thus, unlike the TP pattern described in the preceding section, the development of Rhemes continues in a simple-linear manner, except that their progression beyond the introduction section is interrupted more frequently by links to earlier Rhemes. This is illustrated in Figure 7, representing the RP diagram of the same research article used earlier in the text examples (5–6). The simple-linear shape in the introduction section is a generally unbroken line. Beyond the introduction section, however, while a downward-sloping line is still discernible, we also see a dispersion of Rhemes to the left of this line.

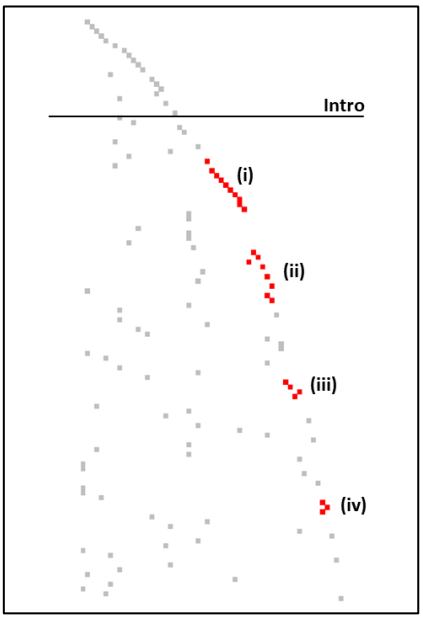


Figure 7. RP showing discontinuous segments

The RP pattern underscores two crucial points about rhematic development. First, new rhematic developments are introduced only *after* the writing has dealt with the previous development.

A visual observation of Figure 7 reveals at least four such developmental clusters; these are marked as (i–iv) and colored in red for easy reference. Each cluster is picked out as comprising three or more Rhemes grouped in proximity to each other. The first two relate to the effects of various nanomaterials on cells, namely, (i) the diffusion of actin, a family of multi-functional proteins, and (ii) the retraction of endothelial cells. These effects and other results then lead the authors to discuss the causes of metastasis in (iii) and the effects on human organs in (iv). This suggests that the rhematic progression of ideas in scientific writing is more than just a simple development of Themes; at the macro level, the picture that emerges is a systematic progression of ideas that pushes the larger message forward.

Second, whereas the RP pattern resembles the TP pattern in the introduction section (in that both are simple-linear in shape), they are distinct in their core functions. As we have seen from (5), the TP in the introduction section represents a *narrowing* of ideas towards the focus of investigation; the RP, by contrast, is

expansionary in its coverage of what the authors did and found. Once the research focus has been established in the introduction section, the rest of the scientific paper then revolves around various aspects related to the focus. The former is achieved thematically, but the latter, rhematically. At the text level, then, the core functions of the Theme and Rheme are not dissimilar to the same at the clause level—the Theme is foundational, and the Rheme is developmental and therefore expansionary.

Third, and perhaps most crucially, the expansionary function of the Rheme captures the very essence of communication. Communicated information, as it were, flows naturally from what is given to what is new (Halliday & Matthiessen, 2014, pp. 116–117). Katz and Odell (2013, p. 94) note that "theories from rhetoric and linguistics support moving from given to new as an effective, successful method of improving communication and learning on a macro level". How this relates to the TP and RP of scientific writing rests on the close relationship between Theme/Rheme and given/new information in an unmarked situation:

> Other things being equal, one information unit is co-extensive with one (ranking) clause ('unmarked tonality'); and, in that case, the ordering of Given ^ New ('unmarked tonicity') means that the Theme falls within the Given, while the New falls within the Rheme. (Halliday & Matthiessen, 2014, p. 120)

At the macro level of the text, this move from established information (the research focus) to new information (the authors' actions and findings) is represented diagrammatically in the TP and RP patterns. The general downward-sloping line in the RP pattern, in particular, underscores the familiar requirement of scholarly publishing, i.e., articles should report on what is novel (or even controversial) about what is known. While the Hallidayan framework does recognize marked instances of information flow, where new information precedes given information, this is impractical in scientific papers since readers will have no way of understanding the new content without any prior contextual information. The skeletal TP and RP patterns, then, reflect the fundamental way in which we communicate. While the scientific paper is indeed about science, it is in essence also about communication.

Conclusion

This study sought to investigate the thematic and rhematic patterns of scientific writing based on a corpus of 50 research articles published in *Nature Nanotechnology* during 2019. The findings are summarized as follows:

(1) TDI was greater than RDI in all contexts—at the whole-text level (3.38 vs. 2.71), in the introduction section (1.82 vs. 1.42), and in the rest of the article (3.19 vs. 2.57). All differences were statistically significant. The articles in the corpus were therefore thematically dense, but rhematically dispersed.

(2) The TP and RP diagrams were different. The TP pattern comprised a simple-linear progression in the introduction section, followed by a constant development in the rest of the article. TDI_{intro} (1.82) was lower than TDI_{rest} (3.19); the difference was statistically significant.

(3) The RP pattern was generally simple linear in its shape. Although RDI_{intro} (1.42) was lower than RDI_{rest} (2.57), the latter was also lower than TDI_{rest} (3.19), accounting for the continuation of a general downward-sloping line, as opposed to the constant development involving Themes. All differences were statistically significant.

In an interesting article on an alternative way for science/professional discourse to be articulated, Rivers (2008, p. 190) argues that "[t]he work of science is communication itself". Yates et al. (2005, p. 36) go one step further, arguing that "[s]cience is fundamentally about communication. Un-communicated science in essence does not exist". The Hallidayan framework captures how the communication of ideas basically works—we first establish a

point of departure (Theme) and then proceed to say more about that point (Rheme). How this is manifested in scientific writing is seen through the TP and RP diagrams, reflecting its basic, skeletal structure. They suggest the Themes in scientific writing orient the reader toward the research focus and anchor the authors as the message points of departure. The actual message itself—i.e., what was done and found—is developed in the Rhemes in a progressive, simple-linear fashion.

Based on the findings in this study, it is conjectured that the structure of scientific writing is layered. On the surface, there is the familiar rhetorical structure, comprising the canonical introduction-methodology-results/discussion sections, or variants of that order. What this study has shown is that there is also a subtler, deeper message structure capturing the very essence of communication. While this message structure may not be as visible as the rhetorical structure, it does ground the writing as, at its very core, an act of communication. Understanding scientific writing this way is helpful in providing both established and novice researchers alike with a broader view of what they (can) do as communicators.

In the light of the paucity of work in this area, the evidence in this study naturally remains tentative. Although the findings regarding the TP pattern corroborate earlier work (Leong et al., 2018; Nwogu & Bloor, 1991), more needs to be done with respect to Rhemes and RP patterns. There is also a need to investigate the robustness of TP and RP patterns using a range of texts from different disciplines in the sciences. Past research on the surface rhetorical structure of scientific writing has revealed differences (e.g., Cargill & O'Connor, 2009), and this is perhaps unsurprising, given the many and varied disciplines in the sciences. Whether the deeper Theme-Rheme structure is more robust, though, remains an open question. If the TP and RP patterns do indeed turn out to be generalizable across the scientific disciplines, we may then have found how scientific writing can be both different (in terms of its rhetorical structure) and similar (in terms of its message structure) at the same time. This is a tantalizing prospect, and more certainly needs to be done to help us more fully understand the communicative aspect of scientific writing.

References

Alley, M. (2018). *The craft of scientific writing* (4 ed.). New York, NY: Springer.

Banks, D. (2017). The extent to which the passive voice is used in the scientific journal article, 1985–2015. *Functional Linguistics*, *4*(1), 12. https://doi.org/10.1186/s40554-017-0045-5

Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science*. Madison, WI: University of Wisconsin Press.

Cargill, M., & O'Connor, P. (2009). *Writing scientific research articles: Strategies and steps*. West Sussex: Wiley-Blackwell.

Costas, R., & van Leeuwen, T. N. (2012). *Approaching the 'Reward Triangle': General analysis of the presence of funding acknowledgements and 'peer interactive communication' in scientific publications*. Centre for Science and Technology Studies (CWTS), Leiden University. https://hdl.handle.net/1887/18648

Cronin, B., Mckenzie, G., & Stiffler, M. (1992). Patterns of acknowledgement. *Journal of Documentation*, *48*(2), 107–122. https://doi.org/10.1108/eb026893

Daneš, F. (1970). One instance of Prague School methodology: Functional analysis of utterance and text. In P. L. Garvin (Ed.), *Method and theory in linguistics* (pp. 132–140). The Hague: Mouton.

Daneš, F. (1974). Functional sentence perspective and the organization of the text. In F. Daneš (Ed.), *Papers on functional sentence perspective* (pp. 106–128). The Hague: Mouton.

Darabad, A. M. (2016). Move analysis of research article abstracts: A cross-disciplinary study. *International Journal of Linguistics*, *8*(2), 125–140. https://doi.org/10.5296/ijl.v8i2.9379

Dimković-Telebaković, G. (2012). Genre analysis: Changes in research article introductions. In H. Sauer & G.

Waxenberger (Eds.), *English historical linguistics 2008 (Volume II: Words, texts and genres)* (pp. 255–266). Amsterdam: John Benjamins.

Ding, D. D. (1998). Rationality reborn: Historical roots of the passive voice in scientific discourse. In J. T. Battalio (Ed.), *Essays in the study of scientific discourse: Methods, practice, and pedagogy* (pp. 117–135). Stamford, CT: Ablex.

Dubois, B. L. (1987). A reformulation of thematic progression typology. *Text*, *7*(2), 89–116. https://doi.org/10.1515/text.1.1987.7.2.89

Enkvist, N. E. (1974). Theme dynamics and style: An experiment. *Studia Anglica Posnaniensia*, *5*, 127–135. http://ifa.amu.edu.pl/sap/files/5/12_enkvist.pdf

Fries, P. H., & Francis, G. (1992). Exploring theme: Problems for research. *Occasional Papers in Systemic Linguistics*, *6*, 45–59.

Gross, A. G., Harmon, J. E., & Reidy, M. (2002). *Communicating science: The scientific article from the 17th century to the present*. Oxford: Oxford University Press.

Halliday, M. A. K. (1967). Notes on transitivity and theme in English (part 2). *Journal of Linguistics*, *3*(2), 199–244. https://doi.org/10.1017/S0022226700016613

Halliday, M. A. K. (1968). Notes on transitivity and theme in English (part 3). *Journal of Linguistics*, *4*(2), 179–215. https://doi.org/10.1017/S0022226700001882

Halliday, M. A. K. (1970). Language structure and language function. In J. Lyons (Ed.), *New horizons in linguistics* (pp. 140–165). Harmondsworth: Penguin Books Ltd.

Halliday, M. A. K. (1994). *An introduction to functional grammar* (2 ed.). London: Arnold.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's introduction to functional grammar* (4 ed.). London: Routledge.

Kanoksilapatham, B. (2005). Rhetorical structure of biochemistry research articles. *English for Specific Purposes*, *24*(3), 269–292. https://doi.org/10.1016/j.esp.2004.08.003

Katz, S. M., & Odell, L. (2013). Something old, something new: Integrating presentation software into the "writing" course. In T. Bowen & C. Whithaus (Eds.), *Multimodal literacies and emerging genres* (pp. 90–110). Pittsburgh, PA: University of Pittsburgh Press.

Khaldoyanidi, A., & Morel, M.-A. (2013). Syntactic properties of the rheme in Russian. *Foreign Language Teaching*, *40*(1), 9–26.

Larivière, V., Haustein, S., & Mongeon, P. (2015). The oligopoly of academic publishers in the digital era. *PLoS ONE*, *10*(6), e0127502. https://doi.org/10.1371/journal.pone.0127502

Leong, P. A. (2014). The passive voice in scientific writing: The current norm in science journals. *Journal of Science Communication*, *13*(1), A03. https://doi.org/10.22323/2.13010203

Leong, P. A. (2015). Topical themes and thematic progression: The 'picture' of research articles. *Text & Talk*, *35*(3), 289–315. https://doi.org/10.1515/text-2015-0001

Leong, P. A. (2016). Thematic density of research-article abstracts: A systemic-functional account. *Word*, *62*(4), 209–227. https://doi.org/10.1080/00437956.2016.1248668

Leong, P. A. (2019). Visualizing texts: A tool for generating thematic-progression diagrams. *Functional Linguistics*, *6*, Article 4. https://doi.org/10.1186/s40554-019-0069-0

Leong, P. A., Toh, A. L. L., & Chin, S. F. (2018). Examining structure in scientific research articles: A study of thematic progression and thematic density. *Written Communication*, *35*(3), 286–314. https://doi.org/10.1177/0741088318767378

Liubov, K. A. (2011). The use of the grammatical category of indefiniteness for the pre-positional subject-rheme identification. *Tomsk State University Journal*, *343*, 27–29. http://journals.tsu.ru//vestnik/en/&journal_page=archive&id=856&article_id=6746

Martínez, I. A. (2005). Native and non-native writers' use of first person pronouns in

the different sections of biology research articles in English. *Journal of Second Language Writing*, *14*(3), 174–190. https://doi.org/10.1016/j.jslw.2005.06.001

McCabe, A. M. (1999). *Theme and thematic patterns in Spanish and English history texts* [Doctoral thesis, Aston University].

Nature. (2020). *How to write your paper*. https://www.nature.com/nature-portfolio/for-authors/write

Nwogu, K. N., & Bloor, T. (1991). Thematic progression in professional and popular medical texts. In E. Ventola (Ed.), *Functional and systemic linguistics: Approaches and uses* (pp. 369–384). Berlin: Walter de Gruyter.

Rivers, N. A. (2008). Some assembly required: The Latourian collective and the banal work of technical and professional communication. *Journal of Technical Writing and Communication*, *38*(3), 189–206. https://doi.org/10.2190/TW.38.3.b

Swales, J. M. (1981). *Aspects of article introductions*. Birmingham: Aston University.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills* (2nd ed.). Ann Arbor, MI: University of Michigan Press.

Udilova, D. (2009). Functions of rheme in the beginning of a sentence in modern Swedish. *Skandinavskaya Filologiya*, *10*, 210–217. http://scandphil.spbu.ru/en/issues/2009-issues/issue-10/functions-of-rheme-in-the-beginning-of-a-sentence-in-modern-swedish/

Viti, C. (2008). Rheme before theme in the noun phrase: A case study from Ancient Greek. *Studies in Language*, *32*(4), 894–915. https://doi.org/10.1075/sl.32.4.05vit

Williams, I. A. (1999). Results sections of medical research articles: Analysis of rhetorical categories for pedagogical purposes. *English for Specific Purposes*, *18*(4), 347–366. https://doi.org/10.1016/S0889-4906(98)00003-9

Williams, I. A. (2009). Discourse style and theme–rheme progression in biomedical research article discussions. *Languages in Contrast*, *9*(2), 225–266. https://doi.org/10.1075/lic.9.2.03wil

Yates, S. J., Williams, N., & Dujardin, A.-F. (2005). Writing geology: Key communication competencies for geoscience. *Planet*, *15*(1), 36–41. https://doi.org/10.11120/plan.2005.00150036

Zaiontz, C. (2020). *Real statistics using Excel*. Retrieved 14 December 2020 from http://www.real-statistics.com/