Chapter 11
Scientific Research Articles:
Twenty-Two Language Errors to Avoid

John Blake
https://orcid.org/0000-0002-3150-4995
University of Aizu, Japan

ABSTRACT

Error-free scientific research articles are more likely to be accepted for publication than those permeated with errors. This chapter identifies, describes, and explains how to avoid 22 common language errors. Scientists need to master the genre of scientific writing to conform to the generic expectations of the community of practice. Based on a systematic analysis of the pedagogic literature, five categories of errors were identified in scientific research articles namely accuracy, brevity, clarity, objectivity, and formality. To gain a more in-depth understanding of the errors, a corpus investigation of scientific articles was conducted. A corpus of 200 draft research articles submitted for internal review at a research institute with university status was compiled, annotated, and analyzed. This investigation showed empirically the types of errors within these categories that may impinge on publication success. In total, 22 specific types of language errors were identified. These errors are explained, and ways for overcoming each of them are described.

INTRODUCTION

English is the *de facto* language of scientific communication (Swales, 1997; Simionescu and Simion, 2004). This means that scientists, regardless of origin and mother tongue, who aspire to publish world-class research need to read and write scientific research articles in English (Ventola, 1994; Bitetti and Ferreras, 2017).

DOI: 10.4018/978-1-7998-4534-8.ch011

Publishing in top-tier English language scientific journals is a prerequisite for career advancement in many countries, or as Ventola (1992) puts it, "publish in English or perish" (p.191). Graduation from doctoral programs, gaining a university post, drawing down research funding and securing tenure may all be contingent on having been published in journals of a pre-determined rank or impact factor (Schein, Farndon and Fingerhut, 2000; Peat, Elliott, Baur and Keena, 2002).

Despite the focus on novelty and substance of research contributions, research articles that are permeated with lexical, grammatical or genre-related errors are more likely to be rejected. Intrusive errors may cause reviewers to misunderstand the intended meaning, severely lowering the chance for acceptance of a manuscript submitted for publication. One non-intrusive language error is unlikely to result in rejection, but multiple non-intrusive errors may create a negative impression possibly leading the academic gatekeepers to fall victim to the horns-and-halo effect, and assume that the research itself may also contain errors (Maiorana and Mayer, 2018). In short, writing error-free prose reduces the possibility of rejection.

This chapter focuses on the language difficulties that scientists need to overcome to write scientific research articles that conform to the generic expectations of the community of practice (Lave and Wenger, 1991). By pinpointing the potential points of failure, i.e. the language errors, scientists can take proactive measures to avoid making similar errors, and adopt a systematic approach to identify these types of errors in their own writing.

Twenty-two specific types of language errors are identified, explained and suggestions for addressing them are given. As there are no native speakers of scientific communication, these errors apply to both native and non-native English speakers. Although this chapter limits its claims to English and scientific writing, it is likely that similar types of errors are present in formal scientific texts written in other languages, and in other formal writing genres, such as academic and research writing.

This chapter starts by reviewing the literature on language errors with particular reference to scientific writing and scientific research articles. However, among the various classification systems that are discussed, no system focusses on the language errors that potentially lead to rejection by reviewers. The following section describes the attempt to address this lacuna in the research by conducting a template analysis of the pedagogic literature. The next section describes the annotation and analysis of a corpus of draft research articles collected at a research institute in Japan. Starting with the five categories identified by the template analysis, the corpus investigation revealed a total of 22 types of errors. The following sections describe, explain and suggest ways to address these errors. The chapter concludes by summarizing the key concepts and identifying future research directions.

The primary objectives of this chapter are to:

- show the importance of error-free writing to scientists and researchers
- demonstrate how genre, practitioner expectations and scientific writing are intertwined
- detail a study to investigate the type of errors in scientific writing
- describe, explain and exemplify the type of errors discovered
- provide a taxonomy of language-related errors
- exemplify and explain each type of error
- suggest ways to avoid or ameliorate these errors

BACKGROUND

Scientific writing is characterized by its lexical complexity, lexical density (Halliday and Martin, 1993; Hyland, 2006) and informational density (Holtz, 2009). The use of compact syntactic structure and the extensive use of nominalization (Halliday & Martin, 1993; Halliday, 1994; Holtz, 2009; Biber and Gray, 2013) result in texts that are dense and often impenetrable to lay readers (Hyland, 2006; Klein *et al.*, 2017). Despite the linguistic and technical complexity of this genre, errors that occur in scientific writing are essentially the same types of errors that occur in academic writing. Common advice from journal editors is to use professional editing services or ask a native English speaker to proofread. Researchers in non-Anglophone countries may not have access to either the funds for professional service or native-speaking colleagues (Canagarajah, 1996). There is also considerable variation in the relative difficulty of learning English with mother tongue being one of the key factors. Languages that are more linguistically distant (i.e. less similar) are thought to be more difficult to acquire (Richards and Schmidt, 2002). Therefore, it may be particularly difficult for scientists whose first language is radically different to English, such as Chinese and Japanese, to write error-free scientific texts.

Errors in writing can be categorized in numerous ways, such as by the source of the error (Fries, 1945; Lado, 1957; Selinker, 1972; Brown, 2000; Wee, 2009), the structure or form of the error (James, 1998; Nesselhauf, 2003), the relative or absolute frequency (Orr and Yamazaki, 2004), the language proficiency of the learner (Selinker, 1972; Grefen, 1979; Edge, 1990) or semantic impact (Burt and Kiparsky,1974; Elliot, 1983).

Source error can be attributed to either a person (e.g. teacher) or language (e.g. mother tongue) that is considered to be the source of the error. The source of the error may stem from the learners themselves (learner-induced) or be the result of the teaching received (teacher-induced) (Ferris, 2006). Intralingual errors (Fries, 1945, Lado, 1957; Selinker, 1972; Brown, 2000) are those that originate in the target language itself while interlingual errors are errors that are ascribed to interference from another language, which in many cases is the mother tongue. Intralingual errors include *inter alia* simplification, overgeneralization, hypercorrection and fossilization (Touchie, 1986).

Edge (1990) extended the works of earlier researchers, such as Gefen (1979) by proposing a classification system that focuses on the relationship between learner proficiency and error type. The slip-error-attempt system advocated by Edge (*ibid.*) can be used by teachers to assess the type of error correction that could be used. Accidental slips are errors that learners can self correct, ingrained errors are those that learners or their peers at the same proficiency level may be aware of while attempts are errors that are beyond the learners' proficiency level, which means that teacher correction is recommended. Ferris (2011) suggests the dichotomy of treatable (rule-based errors) versus untreatable (idiosyncratic) errors.

Errors may also be categorized by the form of written text to which they may be determined to affect. Three common candidate error types are lexical, grammatical and discoursal. Nesselhauf (2003) delineates errors on a scale from morphemic to lexical to syntactic. Lexical errors can themselves be further classified into three categories namely mis-selection, misformation and distortion (James, 1998, p.145). Mis-selection refers to the wrong word choice, misinformation refers to words that exist in the first language but not the target language while distortion refers to words that do not exist in either the first or target languages. Some linguists, particularly systemic functional linguists, may argue that lexical and grammatical should be combined into one category namely lexicogrammatical since grammar itself is realized through words. Orr & Yamazaki (2004) identified twenty errors made in graduation theses drafted by computer science majors at a public university in Japan. The majority of these errors were grammatical located within noun phrases (e.g. articles) or verb phrases (e.g. tense and voice).

Errors that stop the reader understanding the intended meaning or confuse the reader are intrusive (Elliot, 1983) while errors that do not affect the meaning and readers simply ignore are non-intrusive. Although presented as a dichotomy, the reality is a cline from non-intrusive to intrusive. Others researchers (e.g. Burt and Kiparsky,1974; Bates, Lane and Lange,1993) use the term *local* for non-intrusive errors and *global* for intrusive errors.

To assist writers aiming to secure publication, errors that may lead to the rejection of a research article are the focus of attention. These errors are the same types of errors that are likely to result in failing grades should the manuscript be submitted to university faculty. Given the importance of the genre of research communications, many researchers have investigated the reasons that reviewers and editors of journals and conference proceedings reject manuscripts. The four reasons commonly reported as the main causes for rejection are that the research is unoriginal, unimportant, methodologically flawed or plagued by poor language (Bordage, 2001; Pierson, 2004;

McKercher *et al.* 2007; Thrower, 2012; van Weijen, 2014). The first three errors are primarily the concern of the subject specialist while the fourth error, poor language, is the topic of this chapter. Research articles in scientific domains are unlikely to be rejected because of one or two minor language errors. However, intrusive errors may confuse reviewers and stop them understanding the article. Numerous non-intrusive errors may cause negative associations. In short, both intrusive and non-intrusive may reduce the chance that a research article is accepted for publication.

The focus of this chapter is on errors that are more likely to lead to a submission being rejected by reviewers.To identify these critical errors, it is necessary to understand the types or errors that tend to occur in scientific research articles. This knowledge can be used to create a taxonomy that can be used to create checklist for scientists to proofread their manuscripts or serve as the basis for the development of a more complete framework to support writers of scientific research articles. The study was divided into two phases: template analysis and corpus investigation

The starting point for phase one was a survey of the pedagogic literature, which was followed by a template analysis. A detailed description of the procedure and interim results is provided in the template analysis section. The results of the template analysis were then used as the starting point for the subsequent corpus investigation.

PHASE ONE: TEMPLATE ANALYSIS

This investigation was conducted *in situ* at a research institute located on the north-coast of *Honshu*, the largest of the Japanese islands. The postgraduate institution has university status and offers research degrees at masters and doctoral levels. Research is conducted in three disciplines: materials science, information science and knowledge science. The institute is well resourced with library materials in both English and Japanese, and a dedicated writing resource center. The official website claims that the university is bilingual, but interviews with students and faculty revealed that the reality was more bipart-lingual with the Japanese-speaking and the English-speaking communities having limited interaction.

A thorough survey of the pedagogic materials available to support the scientific writing of researchers in this research institute was conducted. Each reference book in the writing section was skimmed to discover whether the contents included information on scientific or research writing. Books that contained at least some reference to scientific research writing were added to the list of resources to be analyzed.

In total, fifteen English language books on scientific communication in the resource center met the criteria for selection for inclusion in the analysis. Each book was read by the author and relevant concepts were coded using template analysis (King, 2004). Template analysis begins with a set of predetermined codes (e.g. grammar, spelling, etc.), but permits *ad hoc* amendment to the initial codes and addition of supplementary codes throughout the study. Template analysis can be viewed as midway between grounded theory with no *a priori* codes and content analysis which prescribes all codes *a priori*.

At the conclusion of the template analysis, the final tagset comprised five categories of errors. The three major categories were accuracy, brevity and clarity; the remaining two minor categories were (perceived) objectivity and formality. Table 1 lists the resource books that were included in the analysis and shows which of the categories in the final five-category tagset were mentioned in each book. Once the template analysis had been completed for the published materials, the highest-ranked websites identified by the Google algorithm for relevant keywords suggested by a focus group of researchers (e.g. scientific writing) were also checked. to see whether their results aligned with those discovered during template analysis.

Although these five categories were identified from the pedagogic literature *per se*, this does not necessarily mean that these categories are reflected in published or draft scientific research articles. The disjuncture between materials designed to teacher English for specific or academic purposes and authentic materials is well documented (e.g. Blake, 2015). In the corpus investigation described in the following section, the applicability of these pedagogic categories is tested and further refined.

PHASE TWO: CORPUS INVESTIGATION

Following the template analysis, a corpus investigation of draft scientific articles was conducted to investigate empirically the types of errors that impinge on publication success. This corpus investigation utilised the findings from the template analysis. The first aim is to ascertain the validity of the results of the template analysis. The garbage-in garbage-out maxim holds. If the content of the pedagogic literature does not reflect reality, then there will be some divergence between the corpus results and the results of the template analysis. The second aim is to extend the granularity of the categories so that teachers of scientific writing may more effectively support novice writers. Frontline teachers of scientific writing who are familiar with the specific types of common errors that writers make can more easily design syllabuses, materials and lessons to help their students avoiding making the same types of errors.

A corpus of two-hundred draft research articles submitted for internal review at the same research institute was compiled. Most authors were PhD candidates, but a small proportion were either faculty or master degree students. Mother tongues of the first authors represented in the corpus include Japanese, Thai, Chinese and Vietnamese.

Table 1. Ethnographic study of the writing resource books published in English in a large research institute in Japan.

Author Surname	Year	Title (abbreviated)	Accuracy	Brevity	Clarity	Objectivity	Formality
Alley	1996	The craft of scientific writing	A	B	C		
Bailey	2014	Academic writing	A		C		
Bentley	2003	Report writing in business	A				
Browner	2012	Publishing and presenting clinical research		B	C		
Fearing & Sparrow	1989	Technical writing: Theory and practice			C	O	
Flowerdew	2014	Academic discourse	A	B	C	O	F
Graves & Graves	2012	A strategic guide to technical communication			C		
Hall	2011	How to write a paper	A	B	C		
Hartley & Buckmann	2001	Business communication	A	B	C		
Laan	2012	The insider's guide to technical writing	A		C		
Ober	2007	Contemporary business communication	A	B	C		
O'Connor	2002	Writing successfully in Science	A		C		
Markel	2012	Technical communication			C		
Matthews & Matthews	2007	Successful scientific writing	A	B	C		
Sageev & Romanowski	2001	A message from recent engineering graduates in the workplace					F
		Totals	10	7	13	2	2

Errors in the corpus were identified manually and semi-automatically using rule-based parsing. On discovering an error, the error was annotated with a tag from the developing tagset. The tagset evolved over the annotation with error types being added and combined until the tagset stabilized. An in-depth description of the methodology of this investigation is available in Blake (2018).

The five pedagogic categories were found to be suitable and functioned well to classify the types of errors in the draft research article corpus. Each of the five categories was further divided into between three and five subcategories, hereafter referred to as types of error. In total, twenty-two types of errors were found in the corpus. The most frequent errors identified were overwhelmingly related to factual or grammatical accuracy. The types of errors are listed below.

Six types of accuracy errors were identified, namely:

1. factual errors in world knowledge,
2. factual errors within the article,

3. overly bold claims,
4. overgeneralization errors,
5. spelling and grammar errors, and
6. statistical and numerical errors.

Errors relating to brevity included:

7. using multiple vague words,
8. repeated words, and
9. redundant words.

There were five types of clarity errors, specifically:

10. use of vague expressions,
11. lexical ambiguity,
12. referential ambiguity,
13. syntactic ambiguity, and
14. garden path sentences.

Errors related to perceived objectivity were categorized into three types:

15. focus on people and feelings rather than things and ideas,
16. emotive wording, and
17. excessive personalization.

The final category of formality errors consists of five types of error, namely:

18. apostrophes,
19. abbreviations,
20. slang,
21. informal terms, and
22. rhetorical questions.

In the following sections, each of these errors is discussed in detail. Concrete examples of the errors have been extracted from the corpus of draft research articles, and where necessary anonymized by substituting specialist terminology with XXX. The extracts contextualize the error description, the error type itself is then described and explained. Practical suggestions are provided to help life-long learners avoid making similar types of errors.

The following five sections detail the twenty-two error types using the error categories of accuracy, brevity, clarity, objectivity and formality. The Pareto principle held for the corpus results with the accuracy category comprising approximately 80% of all the errors identified.

ACCURACY ERRORS

Some of the errors in this category are likely to be considered the most serious by reviewers. The core aim of academic and scientific writing is to convey data, information and/or knowledge with zero distortion of meaning, and so errors that cause degradation of meaning need to be avoided.

Error type 1: Factual errors in world knowledge
Corpus example: The population of Japan is 12,734,100 [1].

There are two main citation styles: author-year and numeric (as in this example). The corpus example is a citation from the source listed as the first item in the reference list indicated by the number in square brackets. There is, however, an error within the citation itself. When copying the number in figures, the writer omitted the final zero, and so the population became 10 times smaller. This type of careless error can cause severe confusion. In this example, readers may either believe the incorrect figure or work out that the figure is incorrect. Through this careless copying error, the writer may have created a negative impression, which could lead the reviewer to succumb to the cognitive bias known as the horn effect and be unduly influenced by this error (MacDougall, Riley, Cameron and McKinstry, 2008). Copying errors are easy to avoid by being systematic. For example, when using the copy-and-paste function, doublecheck that the first and last words or numbers have been included. It is more difficult to make selection errors when using the arrow keys to control the cursor than manipulating a mouse, which relies on fine motor skills, and so the arrow keys are the recommended method of selection. Should it be necessary to copy onto paper, make sure that not only the final numbers or words are included, but that other copying errors do not creep in due to ambiguity caused by poor handwriting. Sixes and zeros may be mistaken for each other, particularly when zeros are not exact ovals.

Error type 2: Factual errors within the article
Corpus example: There are two types of…. First, …. Second, …. Third, ….

When writing a research article, the content of the draft may change significantly. This could be due to changes decided by the writer, but also may be the result of changes suggested or insisted on by co-authors, reviewers, editors or proofreaders. Mistakes in lists are commonplace. The number of items in the list should reflect the number of items that follow. For example, in the list above, initially there were only two items, but apparently a third item was added after the second without changing the introductory sentence which still only lists two items. This type of error appears obvious, but is easily missed by writers. However, people who read this without having seen earlier versions can easily notice such errors.

Error type 3: Overly bold claims
Corpus example: XXX will play a key factor in the near future.

The past has already occurred, the present is occurring now, but the future has yet to come. Although we may attempt to predict the future, we cannot be certain. Even simple events in the future may not happen due to unforeseen circumstances. The modal verb *will* tends to be used to show the certainty of a future action or state. In the example above, the writer expresses certainty. This could become true, but should a natural disaster, war or other event happen, would the claim still hold? Claims that are overly bold can be hedged in three ways. First, the certainty level can be reduced to show a degree of probability, e.g. using modal verbs such as: *could*, *may* or *might*. Second, the scope of the claim can be limited to one that will be certain (or at least no-one could prove otherwise). Third, a condition can be attached to the claim, to add a proviso.

Error type 4: Overgeneralization errors
Corpus example: All women…

Claims about a characteristic of all members of a particular group are problematic. Is it possible to finish the corpus example in a way that is not contentious? Here are some suggestions to consider.

a. All women are not men.
b. All women have two X chromosomes.
c. All women are human.

The first claim, a, presents an either-or dichotomy in which women cannot be men. In logic the law of noncontradiction states that two propositions *A is B* and *A is not B* cannot both be true. However, this law assumes shared definitions of both *A* and *B*. Definitions of women and men vary depending on religious, political or scientific beliefs. Thus, the term *women* can be interpreted in different ways depending on the beliefs related to biological sex and gender, it can be argued that a person born *a* biological male can undertake gender reassignment and become a woman. The second claim, *b*, is also problematic as some people with Swyer syndrome are born with XY chromosomes but have female genitalia and identify as female. The third claim, *c*, is not contentious, since by definition the term women has three elements of meaning: *female, adult, human* and *plural*. When making claims about a groups, it is necessary to play devil's advocate and identify whether it is possible to argue against the claim. If it is, then consider adding modality, limiting the scope or adding a conditional.

Error type 5: Spelling and grammar errors, especially for users of LaTeX
Corpus example: There are three form the first experiment that XXX.

Users of WYSIWYG text processing software, such as Microsoft Word, are used to receiving help from spelling and grammar checkers. Problems can, however, occur when the language settings of the document or sections of the document are not set to English. For example, when pasting text from a non-English document, it is possible to accidently change the language settings, which means that the English words in that section will not be checked. The situation appears even worse for users of LaTeX (Lamport, 1994), which is a plain text editor. This means that writers do not see what their text looks like until it is compiled into a pdf. A key problem for novice uses of LaTeX is not using a spellchecker, since with many free LaTeX editors, spellcheck is not included by default. In the corpus example, Microsoft Word was unable to identify that *form* should have been *from*. Checking spelling and grammar in one's own writing is particularly difficult given the screen-memory interference issue. In short, writers may not necessarily read what is currently written but may read what had been written.

Error type 6: Statistical and numerical errors
Corpus example: (p < 0.5)

The probability value, p, is used to show the probability of finding equivalent or more extreme results while assuming the null hypothesis is true. In short, the smaller the p value the higher the statistical significance. However, reported p values tend to be either $p < 0.05$ or $p < 0.01$. These values are often considered sufficient to claim statistical significance when conducting hypothesis testing. In this example, the mistake was a typographic error that was not noticed in proofreading. The probability of 50% means that the result is random and no significance can be claimed, and the article will most likely be rejected. This careless oversight is similar to the errors discovered in medical records. Hanauer *et al.* (2019) found numerous numerical errors including 1972 invalid dates such as January 35 and February 30. There were also 128 instances of patients aged 150 years old.

BREVITY ERRORS

Error type 7: Using multiple vague words
Corpus example: The concept that was chosen as the focus of this research is XXX.

Genres may be placed on a cline from terse to verbose. Terse genres, such as scientific research articles are concise, but may be considered difficult to read. Verbose genres may be easier to read and understand, but use more words. To quote Faber (2017), “no one wants to read excessively long studies”. Words that do not add substance to the meaning should be omitted. Abstract nouns, such as *concept*, are rather vague. In this corpus example, *concept* refers cataphorically to *XXX*, and as such adds no meaning *per se,* and thus the first six words can be deleted.

Error type 8: Repeated words
Corpus example: We analyze XXX regarding the XXX qualities, XXX qualities and XXX qualities.

Introductory sentences for sections within a research article, tend to list the contents of the section. The corpus example introduces a section that describes three types of qualities. However, the word *qualities* is unnecessarily repeated three times. This sentence can be improved simply by using the word *qualities* once at the head of a noun phrase and then listing the three items as modifiers in the tail of the prepositional phrase that modifies the head. Following this suggestion, the revised concise sentence would be:

We analyze XXX regarding the qualities of XXX, XXX and XXX.

Error type 9: Redundant words
Corpus example: On each and every occurrence, the XXX was noted.

The determiners *each* and *every* have similar but not identical meanings. However, either one on its own is sufficient. They may be used together to show emphasis or to solve translation issues, particularly in legal documents. Legal doublets (Ingels, 2006) are created when near synonyms linked together with a conjunction, which is usually *and*. Cases in point include *aid and abet*, *terms and conditions, goods and* *chattels*, and so on. Although it can be argued, that only one term would suffice, selecting just one near-synonym may result in the target language translation differing in meaning from the source language as the words are near and not exact synonyms. However, in scientific writing, it is unlikely that a researcher needs to stipulate both each occurrence and every occurrence, and so in this case either term would suffice.

CLARITY ERRORS

Error type 10: Use of vague expressions
Corpus example: This is something which is XXX from XXX.

The pronoun *something* is vague as it can refer to so many different items. Clarity in written English is determined by the specificity of the terminology selected. On a scale from clear to obscure, indefinite pronouns, such as *something*, *someone* or *somewhere*, limit their reference to a type of item, namely thing, person or place; but are rather obscure. The clarity of *something* can be increased by substituting a more specific class of item, such as *feature*, *aspect*, *artifact*, *component*, or so forth. Items can be visualized on a cline from vague to clear, as in this example, which shows the cline of obscurity-clarity from *thing* to *toy poodle*.

thing > animal > mammal > dog > poodle > toy poodle

Error type 11: Lexical ambiguity
Corpus example: It is really high for XXX.

In this example *really* is not vague but it has two possible specific meanings. Either it shows the meaning “to a high degree” and can be replaced by the adverb *very* or is used to state the truth or fact of the situation and so could be replaced by the adverb *actually*. However, it is not possible to understand from this decontextualized example (or in fact the example within the original draft research article) what the intended meaning was. This ambiguity can be resolved by replacing *really* with either *actually* or *very* to reflect the intended meaning. In addition, *really* is rather informal, and so fails to meet the genre expectations for the degree of formality.

Error type 12: Referential ambiguity
Corpus example: Referring to Smith [10], Jones notes that he…

The ambiguity is introduced by the third person pronoun *he* which refers anaphorically to an earlier-mentioned person. However, in this context, there are two candidate antecedents, which creates confusion. Referential ambiguity also commonly occurs when pronouns, especially *it* and *this,* are used anaphorically. This type of ambiguity can be resolved by avoiding the need for a pronoun. Thus, the example sentence could be written without the use of a pronoun, thereby alleviating the need to ascertain the antecedent of the pronoun.

a. *Jones notes that Smith [10] ...*
b. *Smith [10] notes that Jones ...*

Error type 13: Syntactic ambiguity
Corpus example: XXX found two AAA and one BBB, which CCC.

The relative clause *which CCC* may modify only one item namely *BBB* or could modify three items namely *two AAA and one BBB*. This ambiguity is caused by using a modifier after noun phrases linked by the conjunction *and*. Resolution is straightforward. One way to limit the modifier to one noun phrase is to bring the noun phrase to the start of a series and place the modifier immediately after that phrase. The other way is when the modifier applies to each item in a series, and so adding a distributive determiner, such as *all* removes any ambiguity. The corpus example could be rewritten as one of the following depending on the intended meaning.

a. *XXX found one BBB, which CCC, and two AAA.*
b. *XXX found two AAA and one BBB, all of which CCC.*

Error type 14: Garden-path sentences
Author-created example: The journal plans to publish your paper were just a rumour.

It was not possible to anonymize the garden-path sentences in the corpus, and so this example sentence was created specifically to exemplify the issue. On reading this sentence, most readers will assume that *plans* is a verb, but on reaching the finite verb *were*, re-read the sentence and re-categorize plans as a *noun*. Garden path sentences are not only ambiguous, but the initially assumed meaning is not the intended meaning. Comedians use this ambiguity for humour (Dynel, 2009), but reviewers find little humour in being confused. Garden path sentences are difficult for writers to notice because the writer of the sentence is primed to focus on the intended meaning. Second readers or proofreaders can, however, easily notice garden path sentences because they need to backtrack to figure out the intended meaning. The difficulty with garden path sentences is not their revision and disambiguation, but their initial identification.

OBJECTIVITY ERRORS

Error type 15: Focus on people and feelings rather than things and ideas
Corpus example: I will describe the results of our research in section 2.

The corpus example uses the first person pronoun *I* and the possessive adjective *our*. This directs attention to the people involved in the research rather than the research itself. This is in contrast to the expectations of the community of practice that focus on things and ideas rather than people and feelings. Research in the many fields within the humanities and social sciences has embraced the use of first person pronouns, *we* and *I,* and no longer eschews active voice in favour of passive voice. However, research in many applied and pure science domains continues to prefer research articles that are appear "objective" in terms of language usage. The corpus example can be depersonalized and written using just five words as follows:

Section 2 describes the results

Error type 16: Emotive wording
Corpus example: We are pleased to announce that the results show XXX.

In the same vein as the previous example, this corpus example does not adhere to disciplinary expectations of "objectivity". The use of the first-person plural pronoun *we* in some scientific domains may be acceptable, but the adjective *pleased* expresses happiness and, as such, focuses on feelings rather than ideas. The author might have been attempting to frame the result announcement in a manner so as to direct the reader to view the results positively. The sentence can be revised by omitted the emotive sentence stem as shown below:

Results show XXX.

Error type 17: Excessive personalization
Corpus example: …such as services to your XXX, to your XXX, and to XXX.

Disciplinary expectations vary among pure and applied sciences, but excessive personalization, such as the overuse of personal pronouns or possessive adjectives is likely to convey the impressive that the research article is not "objective" enough. This "objectivity" refers to the appearance of objectivity through depersonalization of the research narrative. Hyland (2002b) notes that teachers of writing to students with English as an additional language, tend to direct them to depersonalize texts by removing references to themselves from their texts. There is a continuum of usage of first-person *I* in scientific and academic writing. Tang and John (1992) identified six functions realized by the use of first-person *I*, namely: representative, guide, architect, recounter, opinion holder and originator. The opinion holder function seems at odds to the desire to present factual information objectively. However, Hyland (2002a) found that expert writers used I when promoting their own work which was in direct contrast to novice writers who used I to, for example, describe the organization of their research article. The strongest authorial presences created using first person pronouns are when elaborating arguments and stating results or claims (Hyland, 2002a). Perhaps, well-established authors attempt to appeal to their own authority to persuade readers of the validity of their work. However, this is a technique that few would advise novice writers to attempt.

The corpus example can be revised as shown in *a*. This in turn can be made more concise by removing the repeated word, *to*. The final version is shown in *b.*

a. …such as services to XXX, to XXX, and to XXX.
b. *…such as services to XXX, XXX, and XXX.*

FORMALITY ERRORS

Error type 18: Apostrophes
Corpus example: To be more precise, XXX doesn't directly cause the effect (E).

The first type of formality error involves apostrophes. There are two cases when the usage of apostrophes is rather informal. First, when they are used to mark the omission of letters within a word. Second, when apostrophes are used in conjunction with the letter *s* to indicate possession. Revisions are straightforward in both cases. Rather than using contractions, the full form should be used and so *doesn't* should be written as *does not.* Kingdom (2014, p.55) states that "the best approach is to avoid them whenever possible". When apostrophes are used to show possession, the noun phrases can usually be reversed and the conjunction *of* inserted between. For example, *the experiment's rate* becomes *the rate of the experiment.* Another apostrophe problem that affects users of scripts that use different character sets is the type of apostrophe. In the Japanese character set, the apostrophe is close to the western apostrophe, but differs in shape (e.g. ` cf. ').

Error type 19: Abbreviations
Corpus example: This is the RQ of this paper.

The introduction section of many research articles introduces the research question, hypothesis, purpose, aims and/or objective. Students and teachers of research methodology may refer to research questions as RQ, but this abbreviation is neither standard nor common. If the author introduces the abbreviation in brackets after the full form, then the abbreviation can be used later in the research paper. However, unknown or undefined abbreviations should be avoided. Research articles are written from experts in a particular field who share some basic background domain knowledge. This basic knowledge may involve the use of abbreviations. These abbreviations may be acceptable, but it is necessary to check the specific disciplinary discourse as there is significant variation. If there is no advice or precedents to the contrary, the safest option is to avoid using unknown or undefined abbreviations. It should be noted that there is no need to put abbreviations after terms that will not be referred to again later in the research article.

Error type 20: Slang
Corpus example: A bunch of IT engineers collaborated and launched…

Slang describes the type of language used only in specific social circles. These social circles may be related a range of demographic factors, such as age and social status. Slang may creep into scientific writing either through lack of awareness of the linguistic register or poor proofreading. Non-native English speakers who acquire language in Anglophone settings by focussing on improving their spoken communication skills, such as the author of the corpus example; may not realize the inappropriacy of their lexical choices, which in this case is the noun *bunch*. Slang and colloquial language was found by Hanauer et al. (2019) in official medical records. Examples of slang found included *hell of a lot*, *waist deep in* and *boat loads*. Slang errors are easily revised by substituting a neutral or formal term for the offending word. A suggested revision of the corpus example is given below.

Corpus example: A team of IT engineers collaborated and launched...

Error type 21: Informal terms
Corpus example: They launched the website right after the earthquake…

Informal terms differ from slang in that their usage is more widespread and not limited to particular social groups. A common example would be using the word *yeah* instead of *yes*. In the corpus example, the informal adverb *right* is used to mean *immediately*. Informal terms invariably can be replaced with more formal equivalent terms in the same way as slang errors are resolved. The resolution of the corpus example is given in *a.* To avoid any potential referential ambiguity with the plural pronoun *they*, passive voice could be used to rewrite the sentence as in *b*.

a. They launched the website immediately after the earthquake…
b. The website was launched immediately after the earthquake…

Error type 22: Rhetorical questions
Corpus example: The key question to ask is: how can we XXX ?

Rhetorical questions are taught in many high school English classes as a way of interacting with readers. Magazine articles and blogs frequently make use of such devices. Rhetorical questions, however, are less common in scientific research articles, although they are used in editorials. Novice writers would be advised to avoid such techniques, but this is not to say that they are never used. Well established professors do make use of rhetorical questions to lead readers through arguments by encouraging readers to participate in thought experiments and consider options. Rhetorical questions are not commonly used as hooks to attract interest. Rather than posing a question, the author should state the answer to the question. Thus, the corpus example could be reworded as:

XXX can be achieved by XXXX.

FUTURE RESEARCH DIRECTIONS

This research identified a taxonomy of errors that is intended to help scientists draft research articles that are free of language-related errors. The taxonomy provides a theoretical foundation on which resources can be created to support scientific writers.

Of the extant software designed for scientific writing, SWAN, the Scientific Writing AssistaNt is particularly useful (Kinnunen *et al.* 2012). SWAN is a downloadable Java software package that uses a rule-based system to provide feedback to writers using quality metrics. However, there is a niche for an online tool that specifically focuses on providing genre-specific feedback to writers on the language used in scientific texts.

This typology of twenty-two errors served as the theoretical foundation for the creation of a prototype error detection tool (Blake, 2018; Blake, in press) for written scientific communications. The future research direction of this research is the practical application of the taxonomy to help scientific writers. This approach is twofold. First is the refinement of the automatic identification of errors to achieve increased precision using both rule-based parsing and probabilistic parsing. Second is the development of a constructive feedback system harnessing multimodal resources that enables writers to correct the errors identified in an efficient and effective manner.

CONCLUSION

The key outcome of this study was the identification of the list of twenty-two common error types categorized into five broad categories. The categories of accuracy, brevity, clarity, objectivity and formality can serve as filters through which written work can be checked. The twenty-two error type taxonomy can be used to serve as a checklist which can be used to proofread scientific writing for language-related errors. This list is complemented with numerous examples of errors extracted from the corpus of draft research articles.

These error categories were created purely for pedagogic purposes. At times, some of the category boundaries are rather fuzzy, meaning that some errors may be caught by more than one category. For some research purposes, this would be a negative outcome. However, the aim of proofreading is to catch all errors, and so providing more than one opportunity to discover an error, is not necessarily a negative outcome.

The taxonomy also acts as a framework for the development of an automatic error detection tool. This software architecture of the tool was created to find the twenty-two error types. Software developers write functions to address each of the errors. At present, the state-of-the-art research in computational linguistics is not able to address all the types of errors, but progress is being and will continue to be made.

Life-long learners armed with knowledge of common errors made in scientific research articles are better placed to be able to reflect on and identify errors in their own writing.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Alley, M. (1996). *The craft of scientific writing*. Springer. doi:10.1007/978-1-4757-2482-0

Bailey, S. (2014). *Academic writing: A handbook for international students*. Routledge. doi:10.4324/9781315768960

Bates, E., Lane, L., & Lange, J. (1993). *Writing clearly: Responding to ESL compositions*. Heinle & Heinle Publishers.

Bentley, T. (2003). *Report Writing in Business*. Elsevier.

Biber, D., & Gray, B. (2013). Nominalizing the verb phrase in scientific writing. In B. Aarts, J. Close, G. Leech, & S. Wallis (Eds.), *The verb phrase in English: Investigating recent language change with corpora* (pp. 99–132). Cambridge University. doi:10.1017/CBO9781139060998.006

Blake, J. (2018). Corpus-based error detector for Computer Science. In *Proceedings of the Fourth Asia Pacific Corpus Linguistics Conference* (pp.50-54). Takamatsu.

Blake, J. (2020). Genre-specific error detection with multimodal feedback. *RELC Journal, 51*(1), 179-187. https://doi.org/10.1177/0033688219898282

Blake, J. (2015). Prescriptive-descriptive disjuncture: Rhetorical organisation of research abstracts in information science. In F. Formato & A. Hardie (Eds.), *Conference proceedings of 8th International Corpus linguistics Conference* (pp. 377–8). Lancaster: Lancaster University.

Bordage, G. (2001). Reasons reviewers reject and accept manuscripts: The strengths and weaknesses in medical education reports. *Academic Medicine*, *76*(9), 889–896. doi:10.1097/00001888-200109000-00010 PMID:11553504

Brown, H. (2000). *Principles of Language Learning and Teaching*. Prentice Hall.

Browner, W. S. (2012). *Publishing and Presenting Clinical Research* (3rd ed.). Lippincott Williams & Wilkins.

Burt, M., & Kiparsky, C. (1978). Global and local mistakes. In J. Schumann & N. Stenson (Eds.), *New Frontiers in Second Language Learning* (pp. 71–79). Newbury House.

Canagarajah, A. S. (1996). “Nondiscursive” requirements in academic publishing, material resources of periphery scholars, and the politics of knowledge production. *Written Communication*, *13*(4), 435–472. doi:10.1177/0741088396013004001

Di Bitetti, M. S., & Ferreras, J. A. (2017). Publish (in English) or perish: The effect on citation rate of using languages other than English in scientific publications. *Ambio*, *46*(1), 121–127. doi:10.100713280-016-0820-7 PMID:27686730

Duly, H., & Burt, M. (1974). Natural sequences in child second language acquisition. *Language Learning*, *24*, 23–40.

Dynel, M. (2009). *Humorous garden-paths: A pragmatic-cognitive study*. Cambridge Scholars Publishing.

Edge, J. (1990). *Mistakes and correction*. Longman.

Elliott, A. B. (1983). *Errors in English*. National University of Singapore Press.

Faber, J. (2017). Writing scientific manuscripts: Most common mistakes. *Dental Press Journal of Orthodontics*, *22*(5), 113–117. doi:10.1590/2177-6709.22.5.113-117.sar PMID:29160351

Fearing, B. E., & Sparrow, W. K. (Eds.). (1989). *Technical writing: Theory and practice*. Modern Language Association.

Ferris, D. R. (2006). Does error feedback help student writers? New evidence on the short-and long-term effects of written error correction. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 81–104)., doi:10.1017/CBO9781139524742.007

Ferris, D. R. (2011). *Treatment of error in second language student writing*. The University of Michigan Press. doi:10.3998/mpub.2173290

Flowerdew, J. (2014). *Academic Discourse*. Routledge. doi:10.4324/9781315838069

Fries, C. (1945). *Teaching and Learning English as a Foreign Language*. University of Michigan.

Gefen, R. (1979). The analysis of pupil's errors. *English Teachers'. Journal*, *22*, 16–24.

Graves, H., & Graves, R. (2012). *A Strategic Guide to Technical Communication*. Broadview.

Hall, G. M. (2011). *How to Write a Paper* (5th ed.). Wiley – Blackwell.

Halliday, M. A. K. (1994). *An Introduction to Functional Grammar* (2nd ed.). Edward Arnold.

Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: Literacy and discursive power*. The Falmer Press.

Hanauer, D. A., Mei, Q., Vydiswaran, V. V., Singh, K., Landis-Lewis, Z., & Weng, C. (2019). Complexities, variations, and errors of numbering within clinical notes: The potential impact on information extraction and cohort-identification. *BMC Medical Informatics and Decision Making*, *19*(3), 75. doi:10.118612911-019-0784-1 PMID:30944012

Holtz, M. (2009). Nominalization in scientific discourse: A corpus-based study of abstracts and research articles. In *Proceedings of the 5th Corpus Linguistics Conference, #341.* Available at: http://ucrel.lancs.ac.uk/publications/cl2009/

Hyland, K. (2002a). Authority and invisibility: Authorial identity in academic writing. *Journal of Pragmatics*, *34*(8), 1091–1112. doi:10.1016/S0378-2166(02)00035-8

Hyland, K. (2002b). Options of identity in academic writing. *ELT Journal*, *56*(4), 351–358. doi:10.1093/elt/56.4.351

Hyland, K. (2006). *English for Academic Purposes: An Advanced Resource Book.* Routledge. doi:10.4324/9780203006603

Ingels, M. B. (2006). *Legal English communication skills*. Acco.

James, C. (1998). *Errors in language learning and use*. Longman.

King, N. (2004). Using templates in the thematic analysis of text. In C. Cassell & G. Symon (Eds.), *Essential guide to qualitative methods in organizational research* (pp. 256–270). Sage. doi:10.4135/9781446280119.n21

Kingdom, W. (2014). English Grammar and Style: Good writing Practice. *Medical Writing*, *23*(1), 55–57. doi:10.1179/2047480613Z.000000000172

Kinnunen, T., Leisma, H., Machunik, M., Kakkonen, T., & Lebrun, J.-L. (2012). SWAN - Scientific Writing AssistaNt. A Tool for Helping Scholars to Write Reader-Friendly Manuscripts. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, (pp. 20–24). Stroudsburg, PA: Association for Computational Linguistics.

Klein, D., Koltin, O., Peleg, M., Portnoy, I., & Greenbaum, D. (2017). Science and Law Separated by Impenetrable Language Barriers: Overcoming Impediments to Much Needed Interactions. *AJOB Neuroscience*, *8*(1), 37–39. doi:10.1080/21507740.2017.1285831

Laan, K. V. (2012). *The Insider's Guide to Technical Writing*. XML Press.

Lado, R. (1957). *Linguistics across Cultures*. University of Michigan.

Lamport, L. (1994). *LATEX: a document preparation system: user's guide and reference manual*. Addison-wesley.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press. doi:10.1017/CBO9780511815355

MacDougall, M., Riley, S. C., Cameron, H. S., & McKinstry, B. (2008). Halos and Horns in the Assessment of Undergraduate Medical Students: A Consistency-Based Approach. *Journal of Applied Quantitative Methods*, *3*(2), 116–128.

Maiorana, F. A., & Mayer, H. F. (2018). How to avoid common errors in writing scientific manuscripts. *European Journal of Plastic Surgery*, *41*(5), 489–494. doi:10.100700238-018-1418-z

Markel, M. (2012). *Technical Communication* (10th ed.). Bedford/St. Martins.

Matthews, J. R., & Matthews, W. (2007). *Successful scientific writing: A step-by-step guide for the biological and medical sciences* (3rd ed.). Cambridge University Press. doi:10.1017/CBO9780511816185

McKercher, B., Law, R., Weber, K., Song, H., & Hsu, C. (2007). Why referees reject manuscripts. *Journal of Hospitality & Tourism Research (Washington, D.C.)*, *31*(4), 455–470. doi:10.1177/1096348007302355

Nesselhauf, N. (2003). The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics*, *24*(2), 223–242. doi:10.1093/applin/24.2.223

O'Connor, M. (2002). *Writing Successfully in Science*. E. & F.M. Spon. doi:10.4324/9780203478684

Ober, S. (2007). *Contemporary Business Communication* (7th ed.). Houghton Mifflin.

Orr, T., & Yamazaki, A. K. (2004). Twenty problems frequently found in English research papers authored by Japanese researchers. In *International Professional Communication Conference Proceedings* (pp. 23–35). 10.1109/IPCC.2004.1375270

Peat, J., Elliott, E., Baur, L., & Keena, V. (2002). *Scientific Writing: Easy When You Know How*. BMJ Books. doi:10.1002/9781118708019

Pierson, D. J. (2004). The top 10 reasons why manuscripts are not accepted for publication. *Respiratory Care*, *49*(10), 1246–1252. PMID:15447812

Richards, J. C., & Schmidt, R. (2002). *Longman's Language Teaching & Applied Linguistics* (3rd ed.). Pearson Education Limited.

Sageev, P., & Romanowski, C. J. (2001). A Message from Recent Engineering Graduates in the Workplace: Results of a Survey on Technical Communication Skills. *Journal of Engineering Education*, *90*(4), 685–693. doi:10.1002/j.2168-9830.2001.tb00660.x

Schein, M., Farndon, J. R., & Fingerhut, A. (2000). Why should a surgeon publish? *British Journal of Surgery*, *87*(1), 3–5. doi:10.1046/j.1365-2168.2000.01373.x PMID:10606903

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, *10*(3), 209–231.

Simionescu, M., & Simion, E. (2004). Scientific Lingua Franca and National Languages at the Crossroads. In P. Drenth & J. Schroots (Eds.), *ALLEA Biennial Yearbook. Critical Topics in Science and Scholarship* (pp. 129–133). ALLEA. Available at https://allea.org/wp-content/uploads/2016/02/Simionescu_-Lingua_Franca.pdf

Swales, J. M. (1997). English as *Tyrannosaurus rex*. *World Englishes*, *16*(3), 373–382. doi:10.1111/1467-971X.00071

Tang, R., & John, S. (1999). The 'I' in identity: Exploring writer identity in student academic writing through first person pronoun. *English for Specific Purposes*, *18*, 23–39. doi:10.1016/S0889-4906(99)00009-5

Thrower, P. (2012). *Eight reasons I rejected your article*. Elsevier online. Available at: http://www.elsevier.com/connect/8-reasons-i-rejected-your-article

Touchie, H. Y. (1986). Second language learning errors: Their types, causes, and treatment. *JALT Journal, 8*(1), 75-80.

van Weijen, D. (2014). *How to overcome common obstacles to publishing in English*. Elsevier online. Available at: https://www.elsevier.com/authors-update/story/publishing-tips/how-to-overcome-common-obstacles-to-publishing-in-english

Ventola, E. (1992). Writing scientific English: Overcoming intercultural problems. *International Journal of Applied Linguistics*, *2*(2), 191–220. doi:10.1111/j.1473-4192.1992.tb00033.x

Ventola, E. (1994). Abstracts as an object of linguistic study. In S. Čmejrková, F. Daneš, & E. Havlová (Eds.), *Writing vs. Speaking: Language, Text, Discourse, Communication* (pp. 333–352). Gunter Narr.

Wee, R. (2009). Sources of errors: An interplay of interlingual influence and intralingual factors. *European Journal of Soil Science*, *11*(2), 349–359.

ADDITIONAL READING

Gennaro, S. (2016). Mistakes to avoid in scientific writing. *Journal of Nursing Scholarship*, *48*(5), 435.

Lebrun, J. L. (2011). *Scientific writing 2.0: A reader and writer's guide*. World Scientific.

Marina, V., & Snuviškiene, G. (2005). Error analysis of scientific papers written by non-native speakers of English. *Transport*, *20*(6), 274–279.

Provenzale, J. M. (2007). Ten principles to improve the likelihood of publication of a scientific manuscript. *AJR. American Journal of Roentgenology*, *188*(5), 1179–1182.

Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills* (Vol. 1). University of Michigan Press.

KEY TERMS AND DEFINITIONS

Ambiguity: Ambiguity occurs when the meaning of an item can be interpreted in two or more ways.

Discourse Community: A discourse community consists of the people who read, write and/or use a particular set of discourses or text types.

Garden-Path Sentences: Sentences that lead most readers to parse the sentence and discover the incorrect interpretation first.

Generic Errors: Errors that occur when the language used would be considered inappropriate by members of the discourse community of the target genre.

Grammatical Errors: Errors that create expressions that would be considered incorrect according to the rules described in prescriptive grammar books.

Lexical Ambiguity: This type of ambiguity occurs when a term, usually a word, has two or more interpretations.

Lexical Errors: Errors that occur at the level of word choice and can be resolved by replacing the word with a more suitable choice.

Referential Ambiguity: This type of ambiguity occurs when a word or phrase can be interpreted to refer to more than one item.

Syntactic Ambiguity: Syntactic ambiguity or structural ambiguity occurs when a sentence can be interpreted in two or more ways due to the word order used in the sentence.

Vagueness: Vagueness is defined as the lack of precision or specific details.