Towards a Typology of Linguistic and Stylistic Errors in Scientific Abstracts Written by Low-proficiency Doctoral Students in France

Anthony Saber, Audrey Cartron, Claire Kloppmann-Lambert & Céline Louis

Abstract To date, few studies have attempted to formulate typologies of errors by non-native speakers (NNS) in English scientific writing. In this study of 123 doctoral dissertation abstracts written by doctoral students in France, we present a tentative typology of frequent errors that covers issues with general grammar, expert grammar and style. In order to specifically ascertain the errors made by students who experience very significant difficulties, the 123 items of our corpus were chosen after an initial review of 1018 abstracts because they demonstrated low linguistic and stylistic proficiency. The typology of errors was sought in support of an error identification exercise in the *Scientific Writing Assessment Program* (SWAP), an English language certification recently developed at ENS Paris-Saclay.[1] Although some disciplinary variation was seen in the distribution of errors, a convergence towards six major error types (determiners, syntax, tense choice, compound phrases, collocations and lack of clarity) was observed (62.96 % of all errors in geoscience, and 83.89 % in mechanical engineering), suggesting that efforts to mitigate errors should primarily focus on these key issues. Another key finding was that, in contrast with previous studies, traditional grammar issues did not represent the bulk of overall errors (52.78 % in geoscience and only 37.32 % in mechanical engineering), while the overall frequency of stylistic errors was high in both corpora (30.25 % in geoscience, 46.05 % in mechanical engineering), showing the importance of errors in relation to genre-specific style. We propose a metric of error frequency, the *Comprehensive Error Ratio* or CER, to assess the overall quality of abstracts written by non-native speakers of English. In conclusion, we suggest that any typology of errors in ESP/EAP contexts results from a trade-off between seeking descriptive specificity and achieving the specific purposes for which a typology is developed.

Keywords abstracts, English for science, error, France, Geoscience, grammar, mechanical engineering, proficiency, scientific writing, style

[1] The *Scientific Writing Assessment Program* (SWAP) is a not-for-profit English language certification used exclusively for educational purposes at ENS Paris-Saclay.

Zitiervorschlag / Citation:
Saber, Anthony/Cartron, Audrey/Kloppmann-Lambert, Claire/Louis, Céline (2020): "Towards a Typology of Linguistic and Stylistic Errors in Scientific Abstracts Written by Low-proficiency Doctoral Students in France." *Fachsprache. Journal of Professional and Scientific Communication* 42.3–4: 90–114.

1 Introduction

In 2016, École normale supérieure Paris-Saclay (henceforth ENS), a French "Grande École" that trains university lecturers and researchers in 12 disciplines[2], introduced SWAP (*Scientific Writing Assessment Program*), an English language certification devised by the establishment's Department of Languages, into its curriculum as a compulsory element that, among other key items, all students need to validate for graduation. SWAP tests a set of foundation skills in scientific writing (adequately structuring an abstract[3], mastering key scientific phraseology, mitigating language and style errors in the draft version of a scientific article, displaying brevity and concision, discussing scientific results in an appropriate style). The certification comprises a section on error detection, which required a typology of common language and stylistic errors in initial versions of scientific articles. A review of available literature revealed that no comprehensive off-the-shelf typology of errors that would be relevant to the certification was available from previous studies. In 1995, Sionis studied communication strategies in scientific articles written in English by Francophone researchers, and underlined a lack of interest in form, leading to discontinuities in the argumentative process; he also pointed "a lack of familiarity with the discourse conventions of science writing in English" (Sionis 1995: 103) but did not propose a formal typology of errors. In 1996, Birch-Bécaas investigated errors in a corpus of 40 first drafts of medical research articles written by Francophone researchers, and did suggest a very useful typology:

> Forty-two different categories were formed and these were then grouped into larger denominations or 'families' such as determination, tense, cohesion, and word order and finally three global categories, grammar, lexis and spelling (para 14). [...] Grammar represents 75 % of the errors and corrections. 2189 of the 2928 [identified errors] concern grammatical problems. Lexis represents 20 %, or 583 occurrences which were mainly problems of collocation, inappropriate lexical choices, L1 interference and use of compounds. Finally there were 156 spelling errors in the corpus, 5 % of the total (Birch Bécaas: para 16).

However, Birch-Bécaas's study mainly focused on grammatical features rather than on discourse and style, while we believed that a useful typology would cover all types of common errors. Moreover, her study addressed errors made by relatively senior researchers in medicine, while we were interested in foundation scientific writing skills for more junior researchers across many disciplines. Thus, a tentative typology was initially formulated in a top-down fashion by senior scientific writing instructors at ENS, based on their teaching and proofreading experience with doctoral students, but it proved partly unsatisfactory as, after taking

[2] Applied Mathematics, Computer Science, Physics, Chemistry, Biology, Mechanical Engineering, Electrical Engineering, Civil Engineering, Design, Economics, Social Sciences, and English Studies.

[3] Abstracts are increasingly important in present-day research as they play a crucial role in making a case for the research that the authors defend and in convincing readers and reviewers to read the rest of the paper or dissertation. In spite of variation across cultures, times and disciplines (Bondi/Lorés Sanz 2014), regularities of language and structure have been highlighted. According to Swales/Feak (2009: 5), abstracts generally follow a pattern of five rhetorical moves, or communicative stages, associated with recurring lexico-grammatical patterns. For instance, in the result section, writers tend to use *That*-clauses to report relevant results, but prefer to use inanimate subjects (such as *this research shows that*) rather than humans subjects such as *We* or *I* (Swales/Feak 2009: 18).

pilot sessions of the certification, students often reported difficulties with the typology that had been initially proposed. This pilot typology, which supported multiple-choice answers in the section of SWAP devoted to error detection, exclusively focused on stylistic features[4] of scientific writing and included five main categories of errors (verbosity, lack of clarity, word choice or wrong collocation, improper use of evaluative language, and excessive stacking of modifiers), which had been singled out as frequent mistakes necessitating specific attention by ENS instructors, based on their previous knowledge of recurring mistakes in students' papers. Although probably not inaccurate, but perhaps too limited in scope and too general in its typological categorization, this initial classification of errors had been established in a limited time frame which, for institutional reasons, required the swift deployment of SWAP in the establishment's curriculum. Time constraints thus initially precluded a systematic review of errors in a corpus of scientific abstracts written by junior Francophone researchers in support of a more comprehensive typology based on a bottom-up approach. This prompted us to undertake the present study, with a view to avoiding any preconceptions of errors, and to provide ENS students with a clear, comprehensive and accurate typology of key issues in scientific writing.

Errors in EFL/ESL or ESP/EAP contexts have been studied in different ways and from different angles in available literature. Albert et al. (2010) report that the history of "error analysis" as a research domain dates back to the 1970s. Hamilton (2015), in his remarks on the seminal works of "error analysis", points to two major contributions that laid the basis for the discipline: Corder's (1967, 1973) taxonomy of errors and James's (1998) analysis of errors, although Palmer (1917, 1921, 1924) had already recognized the importance of the corrective nature of language courses at the beginning of the 20th century. Norris (1983) attempted to categorize errors made by language learners. In recent research, a productive line of research has been that of feedback on errors and/or mitigation of errors by students (Ferris/Roberts 2001, Truscott 2007, Truscott/Yi-ping Hsu 2008, Storch/Wigglesworth 2010, Ferris 2011, Sampson 2012, Van Beuningen/De Jong/Kuiken 2012, Buckingham/Aktuğ-Ekinci 2017).

Several previous studies have also attempted to establish typologies of errors. Doushaq (1986) proposed a typology of errors made by Jordanian students in academic genres. Payre-Ficout/Chevrot (2004) dicussed errors in the use of the preterit tense in texts written by students at high school and university levels. Luzon (2009) demonstrated that Spanish EFL engineering students showed little awareness of phraseological conventions governing the use of first person plural pronouns in report writing. Adopting a wide-angled approach, Albert et al. (2010) registered errors from a great diversity of document types, authors, disciplines and language levels. Bychkovska/Lee (2017) investigated the most common bundle misuses in L2 student writing by analyzing argumentative papers written by Chinese students, and found that most errors were related to grammar, notably articles and prepositions.

Another productive line of research on error identification and characterization has applied corpus-driven approaches, notably investigating non-native learner corpora. Granger/Tyson (1996) and Bolton/Nelson/Hung (2002) discussed issues related to connector usage in English essay writing. Flowerdew (2000) was interested in referential and pragmatic errors in a learner corpus. Gilquin/Granger/Paquot (2007) also highlighted issues faced by non-native

[4] In addition to SWAP, ENS Paris-Saclay students are also required to take Cambridge's CAE certification, and it was initially assumed that frequent linguistic errors would be mitigated by preparing for this demanding English language certification.

learners when writing academic essays, notably lack of register awareness, phraseological infelicities, and semantic misuse. Gaillat (2013) and Hamilton (2015) focused on the writings of graduate and post-graduate students, using automatic part-of-speech error tagging software that enabled them to highlight salient errors and error frequency. Translation errors in learner corpora have also been studied (Kübler et al. 2016).

However, although a significant number of studies have investigated language error in EFL/ESL teaching contexts, to the best of our knowledge only few surveys have been interested in errors made in scientific abstracts or articles (Sionis 1995, Birch 1996, Birch-Bécaas 1996[5]). As we were interested in defining a typology of errors that would support an error identification exercise in a scientific writing certification specifically designed for Francophone students, we did not attempt to build a large, representative learner corpus that would have allowed to analyze error types and error distribution across written productions by a very large number of non-native speakers with various levels of proficiency. In fact, we would claim that such an approach is not methodologically productive, because it does not integrate the impact of differing language proficiency levels in the analysis of errors. In our view, it is not advisable to review mean levels of errors in a cohort comprising both high-, middle- and low-proficiency writers of the second language, when one's priority is mainly to improve the skills of the latter, as such methods may only give access to mean figures not necessarily significant for those writers who do make a large number of errors. We therefore chose to conduct a qualitative study exclusively focusing on errors made by low-proficiency students, as it was assumed that their abstracts would contain both a wide range and a high number of errors, thereby allowing us to pinpoint the most frequent errors that should be included in a comprehensive and useful typology.

2 Methods

A preliminary review of 1018 abstracts was conducted by a team of four analysts knowledgeable in English scientific writing. The abstracts all belonged to two disciplines (geoscience and mechanical engineering, n = 384 and 634, respectively) and were posted between 2003 and 2017 on www.theses.fr, an official repository of all French doctoral dissertations curated by ABES[6], the French government's agency for bibliographic research. As of 2003, all doctoral candidates enrolled in French universities or higher education establishments have been required to post abstracts both in French and in English on this website after defending their dissertations.[7] Mechanical engineering and geoscience were chosen as the two reference disciplines after initial investigations into various disciplinary fields on www.theses.fr suggested that abstracts written by doctoral students in those fields were more affected by errors than in other disciplines, which was consistent with the aim of this preliminary review, i. e. to identify abstracts displaying very significant linguistic or stylistic issues. It was therefore decided to

5 The references cited here only refer to errors made by French writers, but we acknowledge that other NNS-written academic work has also been covered by previous research, with, however, more emphasis on style and phraseology than on the specific topic of errors.

6 ABES: Agence bibliographique de l'enseignement supérieur.

7 Of note, no official instructions for authors are available from ABES for the writing of abstracts. It is therefore likely that doctoral candidates rely on advice from their academic advisors, or seek stylistic inspiration from previously posted abstracts.

only retain the abstracts displaying multiple errors, whose mitigation could therefore be prioritized for certifying ENS students in scientific writing.

Analysts initially read and rated all the abstracts on a three-tier scale (1 = adequate style and language; 2 = some errors, but acceptable style and language; 3 = significant number of errors in the text). Although empirical, and therefore perhaps not extremely accurate, this method allowed to single out the texts that obviously appeared to be of lower quality, pending further analysis. Abstracts that were given a grade 3 (62 in geoscience, 61 in mechanical engineering) were retained for in-depth error analysis. The resulting corpora (representing 15.8 % of the initial geoscience sample, and 9.6 % of the initial mechanical engineering sample) are described in the table below, and will henceforth be designated as the GC and the MEC.

Table 1: Characteristics of the Geoscience and of the Mechanical Engineering corpora

	Geoscience Corpus (GC)	**Mechanical Engineering Corpus (MEC)**
Total number of words	17 723	13 014
Mean number of words	285.9	213.3
Maximum number of words	506	539
Minimum number of words	115	45
Standard deviation	103.79	88.56

Each abstract was then systematically reviewed for errors in a bottom-up fashion (no pre-established list of errors was used) and annotated as precisely as possible (an example of a fully annotated abstract is available in Appendix 2). Three analysts served as first-line investigators, and their findings were vetted by a senior analyst with 20 years' experience in English scientific writing. All types of errors were considered, i. e. errors related to grammar, style but also terminology and phraseology, thus including both general and specialized features of language. Errors were initially classified into a preliminary typology by each analyst, as abstracts were being reviewed. Categories of errors were then consensually validated to obtain a final typology, which we describe in the results section of this article. In the rare cases where an identified error could possibly be classified under several categories, the four analysts compared their views and reached a consensus on the most adequate categorization.

3 Results

Errors were classified into three main categories: general grammar, expert grammar, and stylistic errors. The sub-categories formulated for each main category of errors were as follows:

General grammar: determiners, genitives, incorrect part of speech, intrusive plurals, missing plurals, prepositions, syntax, subject-verb agreement, tense formation, word order. The typology of grammar errors that we formulated is partly consistent with a previous classification formulated by Birch-Bécaas (1996: para 17), who broke down grammar errors into "determination, tense, cohesion, word order, prepositions, modality, voice and miscellaneous (problems of concord, the use of time expressions such as *for, since* and *during*)".

"Expert grammar" is what we turned our attention to next. It is well known that scientific writing can be characterized by certain specific grammatical and syntactic features, such as a frequent use of passive and impersonal structures (Wilkinson 1992, Rowley-Jolivet 2001), specific "conventionalized" tenses (Hinkel 2004) and a high density of noun structures, which

reflect the expectations of disciplinary discourse communities. Biber/Gray (2010), in particular, showed that academic style in English is highly "compressed", owing to the use of noun phrases with numerous pre- and post-modifiers such as adjective phrases, nouns or prepositional phrases. Therefore, while not incorrect from a linguistic point of view, some grammatical choices made by junior Francophone researchers may be perceived as highly inadequate in the context of English scientific writing. These register-specific choices were termed "expert grammar" by Halliday/Martin (1993). From the analysis of the corpus, two sub-categories of expert grammar errors clearly emerged: issues with modifiers and compound phrases, and ill-advised choice of tenses in the context of scientific writing.

Efforts were also made to minimize possible error overlap in the analysis of grammatical issues, notably in the case of an inadequate order of elements in compound phrases (as in "in two experimental devices incubations" instead of "incubation devices" in the GC[8]), which could also fall under the "word order" category: it was decided that all word order issues associated with compound modifiers would fall under the "modifiers and compounds" category, and would therefore pertain to expert grammar issues.

A third category was retained after analyzing the corpus, and was designated as **"stylistic errors"** including four sub-categories: poor choice of collocations or phraseology, complex errors, use of informal or inadequate register, and obscure formulations.

Of note, a fourth category of errors was found, that of loan translations. However, as it represented only 1.54 % and 2.39 % of the overall number of errors in the GC and the MEC (respectively), this type of error was deemed relatively anecdotal.

3.1 Grammar errors

Several general grammar[9] errors were observed in our corpus. The ten sub-categories are shown in Table 2 and Table 3:

Table 2: Analysis of general grammar errors in the GC

	% of overall grammar errors	Mean error ratio (% of total number of words)	Standard deviation of mean error ratios
Determiners	29.94	0.81	0.0060
Syntax	15.80	0.43	0.0052
Prepositions	13.72	0.37	0.0038
Word order	12.06	0.33	0.0050
Subject-verb agreement	7.90	0.21	0.0033
Missing plurals	7.28	0.20	0.0029

[8] Source: http://www.theses.fr/2016LORR0148, accessed in December 2019.

[9] By "general grammar", we mean grammatical choices that do not reflect certain constraints and established traditions of scientific writing in English, e. g. using action verbs in the past tense and the passive voice in the "methods" section of a research article. General grammar can thus be contrasted with "expert grammar" (Halliday/Martin 1993), i. e. grammatical choices that are directly governed by the specialized genre to which the text belongs and that reflect the expectations of a given discourse community.

	% of overall grammar errors	Mean error ratio (% of total number of words)	Standard deviation of mean error ratios
Incorrect part of speech	7.07	0.19	0.0037
Intrusive plurals	6.03	0.16	0.0033
Tense formation	4.99	0.14	0.0039
Genitives	2.29	0.06	0.0020

Table 3: Analysis of general grammar errors in the MEC

	% of overall grammar errors	Mean error ratio (% of total number of words)	Standard deviation of mean error ratios
Determiners	33.63	0.86	0.0062
Syntax	31.23	0.80	0.0083
Prepositions	8.41	0.22	0.0036
Incorrect part of speech	7.81	0.20	0.0046
Word order	7.21	0.18	0.0030
Intrusive plurals	6.01	0.15	0.0036
Tense formation	4.50	0.12	0.0027
Missing plurals	4.20	0.11	0.0028
Subject-verb agreement	2.70	0.07	0.0026
Genitives	2.10	0.05	0.0018

Overall, three types of grammar errors (determiners, syntax and prepositions) accounted for a significant majority of all errors (59.46 % in the GC, 73.27 % in the MEC) in both corpora. Interestingly, syntactic issues were much more present in the MEC than in the GC (31.23 % v. 15.80 %), exemplifying some degree of disciplinary variation in errors, but convergence between the two corpora was also observed for other error types (notably determiners, with relatively close error rates in both corpora, 29.94 % in the GC v. 33.63 % in the MEC). Issues with parts of speech or intrusive plurals also had very similar mean ratios in both corpora. Moreover, the relatively high levels seen in the standard deviation of mean error ratios for those three very frequent types of errors also suggest that error types occur in an uneven fashion in each individual abstract (a pedagogical repercussion of this finding is probably that the mitigation of main grammatical errors should be specifically tailored to each doctoral student).

Incorrect use of determiners

In some cases, given the trend towards brevity and concision of scientific writing, a formulation without a determiner can be considered as acceptable in an abstract or a scientific article, where it would have been expected in "general" English, in compliance with traditional grammar rules, as shown in the example below:

(1) MEC: A method for **systematic synthesis of** spherical parallel mechanisms architectures, based on algebraic proprieties of the Lie group of displacements, was developed.

In example (1), there is no definite or indefinite article before "systematic synthesis". While it might appear as an erroneous choice under traditional grammar rules, the definite article can probably be left out here, owing to the need for efficiency and brevity often observed in scientific writing. However, we also observed more instances in which authors omitted to use the definite article while its use was required, as in "the aircraft industry" in example (2) below (the phrase also has an additional collocational issue, as "aviation industry" would obviously have been a better choice):

(2) MEC: Working context is in the machining without lubrication of the alloy AA on 2017 fluently used in **aircraft industry**.

Conversely, generic plurals were at times used with the definite article, while it should have been left out:

(3) MEC: **The capillary pumped loops** are passive and modular heat transfer devices, characterized by their highly efficient and highly reliable behavior.

In example (3), the article "the" is not needed: using the definite article would imply a specific reference to a certain group of capillary pumped loops already mentioned in the text of the abstract, while the sentence actually provides a generic definition (each and every capillary pump loop in the study can be characterized as a "passive and modular heat transfer device"). On the contrary, example (4) presents a context in which the article is needed. The use of a defining prepositional clause "of the geodynamic evolution" requires the use of a definite article before "main stages":

(4) GC: They are correlated with **main stages of** the geodynamic evolution of the basin.

Issues with the indefinite article were also noted. In example (5), the students misused "an" instead of the article "a":

(5) GC: That methodology is illustrated and applied **to an modelling process**.

Our findings suggest that the proper use of both the definite and indefinite article is not mastered by low-proficiency doctoral students in France, although most of them have received 11 years of formal instruction in English (7 years in secondary school, and generally 4 years at university[10]).

Some errors related to other determiners, especially possessive pronouns, were also observed, as in example (6):

(6) GC: The current calculations have been done to complete **his results**.

The choice of the incorrect possessive "his" (instead of the determiner "these") can probably be explained by the student's mistake regarding the two homophones *ses* ('his') and *ces* ('these') in French. Incorrect use of determiners was also associated with issues related to grammatical agreement, as in example (7):

[10] Under French educational regulations, all Master's degrees at French universities should theoretically offer formal instruction in foreign languages, notably English.

(7) GC: **This images** are greatly improved …

Overall, our findings suggest that determiners, and most notably articles, are very frequently misused by low-proficiency doctoral students in France. Two main reasons may explain this phenomenon: firstly, there are very significant differences in the use of both the indefinite and the definite articles in French and in English, with a much higher number of instances where the 'zero article' is needed in the latter; therefore, L1 interference probably accounts for a significant number of errors with articles. Secondly, formal instruction in English grammar has dramatically declined in the past 30 years in the French secondary school system (which favors learner-centered curricula and a pedagogy of English based on socialization through language[11]).

Incorrect syntactic structures

A large number of syntactic inadequacies were observed in our corpus. Very often, punctuation or intrusive elements disrupted syntactic units (for instance, the verb/object unit, the noun phrase or prepositional phrase unit ...), as in example (8):

(8) GC: Hydrothermal systems could have constitute [sic] a favorable environment for the appearing of life because these systems are characterized by black smokers and white smokers with an ecosystems [sic] independent of solar energize **where the most primitives organisms live, found in actual environments, the hyperthermophiles**.

Another frequent case is the use of run-on sentences, with lack of punctuation, coordinating or subordinating elements, as in example (9):

(9) GC: During this study I developed a post-processing technique **called WAM** (Weighted Average Model) **allows the reconstruction of reliable velocity models of P and S waves**.

In a few cases, we even found incomplete sentence fragments:

(10) GC: During the Neogene a roll-over system developed on the slope of the DBGM**, this roll-over detachment at a surface of clays of the Eocene-Oligocene**.

At times, some sentences with a poor syntax were even barely understandable upon first reading, as in the following example:

(11) GC: The aim of this thesis is to study the sulfur behavior in hydrothermal fluids subjected to low grade metamorphism. **It** is separated into three interconnected studies linking natural and experimental study.

In example (11), the deictic "it" is too far away from its referent, which results in a pronoun whose referent is unclear (one may indeed think that "it" refers to "metamorphism").

[11] This is termed *pédagogie co-actionnelle* in the French secondary school system.

Prepositions

The choice of appropriate prepositions also seemed to be one of the major difficulties faced by low-proficiency doctoral students, as can be seen in example (12):

(12) MEC: Using mesoscopic model and the intrinsic failure criteria, a cohesive model (macroscopic model) in agreement **of** time calculation limitations is identified.

We hypothesize that errors with prepositions were often due to L1 interference. It can be assumed that authors tend to write or at least mentally pre-formulate their abstracts in French first, then translate their texts into English, which may cause significant issues with several grammatical structures, including prepositions. This can be observed in example (13), which erroneously transposes the French expression *pour cela* into English as "for it" (instead of a preferable formulation such as "in order to do so"):

(13) MEC: The final aim is to provide a model able to modelling multi-materials structure under crash loading. **For it,** a characterisation of a phenomenological model (mesoscopic scale) is proposed.

Having described the three main grammatical issues that were identified, we now turn to some selected examples of less frequent errors.

Missing plurals

Some plurals were omitted or missing. Although not very frequent, this error revealed a poor command of plurals indicative of general notions or categories, as in example (14):

(14) GC: The last point is the feasibility of the determination of **rare earth element** in **fluid inclusion**.

One would expect "elements" and "inclusions", as the sentence does not refer to any specific context and is thus endowed with a general scope. A probable explanation of these errors lies in the fact that the French language, contrary to English, generally expresses generic notions in the singular. Therefore, it is very likely that students applied this rule while writing or translating their abstracts into English, and were not aware of differences between French and English in this respect.

Subject-verb agreement

Some errors in subject-verb agreement appeared to reflect a lack of basic grammatical knowledge, notably as regards subject-verb agreement:

(15) MEC: The considered models are widely used for engineering applications and **focus have been made** on the Simo model implemented in finite element commercial software Abaqus.

Tense formation

Example (16) exemplifies errors related to tense formation, especially regarding the formation of the infinitive and of past participles:

(16) GC: This **to better understood** the role of sulfur speciation in the interactions between an aqueous phase and a hydrocarbons fluid during the TSR.

Word order

An incorrect order of components in clauses appeared to be one of the most frequent grammar errors in our corpus. Adjectives (example (17)), adverbs (example (18)), direct objects (example (19)) were misplaced by authors:

(17) GC: This **study natural** was completed by a preliminary experimental study.

(18) GC: I show that the seasonal subsidence **depends greatly** on the geological land unit ...

(19) GC: This technique allows **to heat up to 300°C samples** (natural and synthetic).

Genitives

Incorrect uses of genitives were also observed in our corpus (although at low frequencies, with a mean ratio of 0.06 % in the GC and 0.05 % in the MEC), as in example (20):

(20) MEC: Some comparisons with the results from **the Cauchy's** similarity law highlight a significant improvement in term of the accuracy of the predictions of the plateaus of force.

3.2 Expert grammar errors

As shown in Table 4, our investigations highlighted two major issues in this category of errors: low-proficiency doctoral students appear to struggle with modifiers and compound phrases, and also tend to show poor judgment in tense choice.

Table 4: Analysis of expert grammar errors in the GC

	% of expert grammar errors	**Mean error ratio**	**Standard deviation of mean error ratios**
Modifiers and compound phrases	45.33	0.38	0.0059
Choice of tense	54.67	0.46	0.0059

Table 5: Analysis of expert grammar errors in the MEC

	% of expert grammar errors	**Mean error ratio**	**Standard deviation of mean error ratios**
Modifiers and compound phrases	53.76	0.33	0.0077
Choice of tense	46.24	0.66	0.0073

Modifiers and compound phrases

As shown in Tables 4 and 5, a source of error in terms of "expert grammar" lies in the formation and use of compound nouns and noun phrases. As Birch-Bécaas (1996: para 22) indicated, Francophone researchers need "to incorporate a mass of information in as concise a text as possible. This leads to dense lexis and complex compound constructions […]. For the NNS however, the use and over-use of nouns as pre-modifiers is problematic."

First, we encountered many cases in which the stacking of modifiers was excessive:

(21) GC: It conveys the doubt linked with **our subsoil properties understanding** on probabilistic models, and proposes to integrate it on them.

In addition, the formation of compounds seemed to be an important source of error. Several errors involved the use of an intrusive -s ending in the modifier noun, while this is generally forbidden by grammar rules (with a few exceptions such as "arms race"), as in example (22):

(22) GC: **Tools development** for geophysical data interpretation …

As already mentioned above, a third category of errors in compound modifiers involved incorrect component order, as in example (23), where the correct order of words in the compound phrase would probably be "the study of hot steel brittleness":

(23) MEC: The aim of this thesis is to optimize a thermo-mechanical test dedicate[d] to the study of **a hot brittleness** of steels.

Even if, as Mignot (2001: 2) points out, compounding is a very productive way to form nouns in English, researchers for whom English is their second language thus need to be careful in the formulation of new compounds. Indeed, Biber/Gray (2010: 11) showed that in a "compressed noun phrase" (which are frequently used in English language scientific writing) it can be difficult to ascertain the accurate semantic relationship between the head noun and phrasal modifiers.

Tense choice

Among the most common errors in our corpus were errors in the choice of tenses, in relation to expert grammar. Hinkel (2004) demonstrated that even after many years of L2 instruction and practice, advanced NNS students still experience difficulty with the conventionalized uses of tenses, aspects and the passive voice in written academic discourse. As Birch-Bécaas indicates, "The French researcher […] tends to confuse the preterit and the present perfect for the description of procedures" (1996: para 20). Indeed, it is the simple past that is chiefly used in academic writing (in the active or passive form) to describe the different steps in the study, notably methods and results. In breach of this established convention, and consistent with Birch-Bécaas findings, many authors in our corpus prefer to use the present perfect, particularly while referring to their own methodology, as in example (24), although the use of the present perfect should be limited to reference to past studies and results obtained by other researchers (in a review of available literature on the subject, as in the phrase "previous studies have shown that …", for example):

(24) MEC: The tribological and spectrometric measurements with the rubbed surfaces analysis **have permitted** to distinguish two varieties of tribological behaviour.

Additionally, research articles, dissertations and abstracts use the simple present almost exclusively to express general truths, such as the background of the study, or generalizable conclusions. Some Francophone authors, conversely, seem to be tempted to use the present to describe what they did to obtain their results, instead of opting for the simple past, thereby demonstrating poor proficiency in the conventionalized use of tenses, as in example (25):

(25) GC: **In Tibet, I process data** from the Envisat satellite archives, at the boundary of the Tibetan plateau, in two seismic gaps, **which appear** interesting to study the partitioning of the convergence.

3.3 Stylistic errors

A significant number of errors (30.26 % and 46.05 % of the overall total of errors in the GC and the MEC, respectively) had to do with style in our corpus. Stylistic errors can affect the clarity and readability of an abstract, or transgress rules of academic writing implicitly or explicitly imposed by discourse communities. Very long sentences, for example, are not advised in scientific writing, because they may lead to considerable lack of clarity, as in example (26):

(26) GC: In this way**, the Curiosity rover that travels** in Gale crater, **which formed** by impact during the Hesperian period (3.5-3.8 Gyr) within igneous basement rocks dated at 4.2 Gyr, **discovered** Noachian alkaline igneous rocks (> 3.8 Gyr) using the ChemCam LIBS instrument ('laser induced breakdown spectroscopy').

As can be seen in the example, embedded relative clauses ("that travels …", "which formed", "dated at 4.2") hinder the understanding of the structure of the sentence, especially that of the matrix clause "the curiosity rover […] discovered Noachian alkaline rocks (> 3.8 Gyr)".

However, we observed very few overly long sentences in our corpus (the fact that we investigated abstracts, i. e. relatively short texts generally comprising 200 to 300 words, puts this finding in perspective). Therefore, we did not retain this first type of stylistic error in our typology (although it may be a significant error in full-length research articles written by low-proficiency students), but other significant errors were observed, as shown in Tables 6 and 7:

Table 6: Analysis of stylistic errors in the GC

	% of stylistic errors	**Mean error ratio**	**Standard deviation of error ratios**
Collocations and phraseology	54.58	0.91	0.0079
Complex errors	11.86	0.20	0.0034
Informal or inadequate register	5.42	0.09	0.0028
Obscure formulations	28.14	0.47	0.0079

Table 7: Analysis of stylistic errors in the MEC

	% of stylistic errors	Mean error ratio	Standard deviation of error ratios
Collocations and phraseology	56.21	0.81	0.0124
Complex errors	4.74	0.07	0.0044
Informal or inadequate register	3.84	0.06	0.0027
Obscure formulations	35.21	0.051	0.0107

Interestingly, the standard deviations of stylistic error ratios revealed relatively high variability in the incidence of issues with collocations and obscure formulations in both corpora, whereas complex errors or register issues were more evenly distributed. This finding probably reflects the idiosyncrasies of individual writers when formulating their ideas, but warrants further study (structured interviews of authors would certainly be useful to ascertain the reasons why they made some incorrect formulation or stylistic choices).

Collocations and phraseology

We first grouped together all the errors that appeared to be related to inadequate word choice or phraseology. Almost all collocational issues in the GC and the MEC had to do with established conventions of Anglophone scientific writing, which has a certain phraseological fixity, as amply demonstrated in previous studies (Gledhill 2000a, 2000b, Hyland 2008, Cortes 2013, Pan/Reppen/Biber 2016, to quote but a few surveys). Some words or some formulations, notably those used as metalanguage to describe the inner workings of the research itself, may be ill-chosen for a scientific paper, such as "in this work" in example (27) is used instead of, among a range of possible adequate formulations, "in this study".

(27) GC: **In this work**, a many experimental campaign was carried out.

However, most of the errors were not only due to the use of an isolated lexical item, but rather to that of a complete lexical sequence, which would probably be perceived as inadequate in terms of scientific phraseology by proficient scientific writers or by reviewers. In example (28), we would have expected a verb associated with investigations ("determine", "investigate", for example) instead of "concerns":

(28) GC: **The objective of my thesis concerns** the acoustic and seismic properties of unconsolidated formations.

Complex errors

Several passages contained a combination of interrelated multiple errors. Such errors were labelled as "complex errors". Indeed, some segments displayed multi-layered errors that we were unable to classify under a single, specific type. Example (29) illustrates the complex mechanisms at work in those errors:

(29) MEC: **A discussion is made with concern to** experimental set-up as well as the used procedures for an efficient specimens testings.

Here, the overall formulation seems heavy and inadequate: first, the lexical-grammatical pattern "to make a discussion" is inadequate, while the verb "to discuss" would be more suitable. As we know that in scientific academic writing the passive voice (*x is discussed*) is at times more appropriate than the active one (*we discuss*), especially when referring to methods or objectives, our suggested formulation would be "the experimental set-up is discussed, as well as …". We would also leave out the heavy and unnecessary phrase "with concern to".

Poor syntax very often compounded problems of other nature in our corpus. In example (30) for instance, both syntax and the choice of terms are problematic:

(30) GC: frictional damping […] is a **parameter difficult to tune**.

The lexical-grammatical pattern "to tune a parameter" does not seem to be adequate here, as "tuning" suggests a very fine mastery and control of the parameter with the help of an instrument, which is not what the author means, when we look at the French version of the abstract.

Informal or inadequate register

Other stylistic errors included errors of register, such as informal formulations, or contracted forms, which are not usually found in academic writing.

(31) MEC: Consequently, **it's** pertinent to make these studies again with the modern means of simulation.

Furthermore, inadequate evaluative language was found in our corpus. In example (32) for instance, the evaluative adjective "simplistic" could be read as slightly offensive to researchers who suggested the interpretation quoted by the author:

(32) GC: Until recently, Mars was considered as a planet with a homogeneous crust dominated by olivine-rich basalts. This **simplistic** vision has been largely disrupted especially with results of recent in situ missions.

In this example, hedging could also have been useful to show more respect for alternative visions: "this vision may be incomplete" could have replaced the phrase in bold. Hyland (2008: 102) suggested that the avoidance of categorical assertions shows both acceptance of alternative views and recognition that one's judgment or assertion is subjective and/or temporary. It seems clear that the author of example (32) was not aware of those Hylandian lexico-grammatical codes.

Obscure formulations

Other stylistic errors were categorized as obscure formulations when it was difficult for the analyst to ascertain or even to guess the meaning of the phrase. For instance, example (33) shows how the very message of the abstract can be affected by obscure structures:

(33) GC: **This opening has the relative movement towards the southeast of the Yucatan Block, this will be the origin of the DBGM** […].

The main problem of the sentence lies with the semantics of the verb "has": did the author intend to say that "the opening" *triggers* – and not "has"– "the relative movement towards the southeast of the Yucatan Block"? Or, as the French translation suggests, is there a confusion

between the verb "has" and coordinating conjunctions "and" or "as well as", in which case the syntax of the whole segment needs to be revised ("the opening as well as the relative movement towards the Yucatan Block is at the origin of the DBGM")?

4 Discussion

Establishing a typology of errors in scientific writing is a rather arduous task, as the choice of categories may seem arbitrary at times, or may categorize errors in a heterogeneous fashion (errors may affect grammar, register, style, text structure …), with different levels of focus (errors may be analyzed at part-of-speech level, sentence level, paragraph level, text level, or even in relation to pragmatic or cultural situations and contexts). Thematic-rhematic structure may also serve to construe errors, notably in scientific articles, where informational patterning can be highly codified (Carter-Thomas 1998). In this study, our bottom-up approach led us to formulate a heterogeneous typology that included both grammatical and stylistic features, but other approaches could of course be relevant. Besides, errors may at times be difficult to categorize, or may be intermingled to form "complex errors", as suggested above.

Our bottom-up approach allowed us to formulate error types "on the go", as errors were gradually discovered in the corpus, but it also factored in the purposes for which we undertook this study. Indeed, in our view, a successful typology of errors should be neither too broad (e. g. with very generic categories such as "lexis", "grammar", etc.), nor overly specific (e. g. with multiple, very narrow categories such as, hypothetically, "improper order of components in cleft clauses"), notably when the projected typology will be used in support of ESP/EAP instruction modules and/or for certification purposes, as was the case in this study. In other words, as our ultimate objective was to include the typology of errors in a multiple-choice error detection exercise, we tried to obtain a "workable" classification that offered, in our view, a good tradeoff between descriptive specificity and efficiency for scientific writing teaching and certification purposes. A simple typology based on very general categories, similar to Doushaq's (1986), who distinguished issues at sentence level, paragraph and content level in academic compositions written by Jordanian students, would not have served our purposes. Conversely, a very specific typology leading to several dozen categories would not have been workable for inclusion in a multiple-choice exercise. Hence, we decided, for example, to use a broad, encompassing category ("syntax") for syntactic errors or inconsistencies, although, of course, such errors could have been broken down into multiple sub-categories (problems with subject-predicate relation, with complements, with adjuncts, with prepositional phrases, with subordinate clauses, etc.).

Although some degree of dispersion and variability were observed in the distribution of error types amongst the GC and the MEC, our typology did highlight some very frequent error types, as shown in Table 8:

Table 8: Overall breakdown of error types

	GEOSCIENCE % of overall errors		MECH. ENGINEERING % of overall errors
Collocations and phraseology	**16.51**	**Collocations and phraseology**	**25.88**
Determiners	**14.77**	**Obscure formulations**	**16.22**
Obscure formulations	**8.51**	**Determiners**	**11.64**
Tense choice	**8.41**	**Syntax**	**10.81**
Syntax	**7.79**	**Modifiers and compound phrases**	**10.40**
Modifiers and compound phrases	**6.97**	**Tense choice**	**8.94**
Prepositions	6.77	Prepositions	2.91
Word order	5.95	Incorrect part of speech	2.70
Subject-verb agreement	3.90	Word order	2.49
Complex errors	3.59	Loan translations	2.39
Missing plurals	3.59	Complex errors	2.18
Incorrect part of speech	3.49	Intrusive plurals	2.08
Intrusive plurals	2.97	Informal or inadequate register	1.77
Tense formation	2.46	Tense formation	1.56
Informal or inadequate register	1.64	Missing plurals	1.46
Loan translations	1.54	Subject-verb agreement	0.94
Genitives	1.13	Genitives	0.73

As can be seen in Table 8, six priorities in terms of error mitigation clearly emerge from our typology: collocations and phraseology, determiners, modifiers and compound phrases, obscure formulations, syntax, tense choice represented 62.96 % of all errors in the GC, and 83.89 % in the MEC. A seventh type of error (issues with prepositions) also appeared at a relatively high frequency (6.77 %) in the GC, while appearing as more anecdotal (2.91 %) in the MEC. It can therefore be hypothesized that, instead of trying to mitigate all errors in a comprehensive fashion, scientific writing instructors could gain significant efficiency by concentrating most pedagogical efforts on these very frequent errors in both corpora, and more specifically on adequate collocations and stylistic clarity (rather than on more traditional grammar issues). Indeed, the overall proportion of traditional grammar issues (excluding "expert grammar" errors) was 52.78 % in the GC and only 37.32 % in the MEC, to be compared with the overall frequency of stylistic errors (30.25 % in the GC, 46.05 % in the MEC). Our finding that traditional grammar issues did not represent the bulk of overall errors, notably in the MEC, is in

contrast with the results obtained by Birch-Bécaas in 1996 and with Bychkovska/Lee in 2017, who found a significant preponderance of grammatical issues in the errors they analyzed.

It also seems necessary to factor in possible disciplinary variation: for example, Table 8 shows that there were twice as many obscure formulations in the MEC compared to the GC (16.22 % v. 8.51 %), which warrants further investigations. While assessing the impact of disciplinary variation was not a major outcome of this study, we hypothesize that the higher number of issues with prepositions in the GC may have been due to the need to describe the spatial characteristics of geological terrains or features – as in

(34) GC: "These results explain complex rift systems with both vertical penetration of plume material **into** the overlying lithosphere"

– thereby necessitating a more frequent use of prepositions.

Determining the causes of errors was not the purpose of this study, but we can hypothesize that they are due to multiple factors, notably, as already discussed above, L1 interference (although, interestingly, the number of direct loan translations was low in both corpora, a finding that is perhaps at odds with traditional beliefs in English language instructors, the influence of French could clearly be seen in determiners, syntactic structure, collocational choices and compound modifiers). It is likely that most authors wrote their abstracts in French first, before translating them into English, but we would need to verify this hypothesis through questionnaires or interviews. Other sources of errors may include lack of adequate proofreading, false beliefs (such as the frequent belief that modifiers can be stacked without any limit in English), or even by the improper use of dictionaries. A more detailed study may show that L1 interference is indeed a frequent source of errors, as in example (35):

(35) GC: The use of two frequencies (50 and 500 MHz) as well as comparison with trenches bring **complementaries informations** on the geometry.

The misleading influence of French grammar can be seen twice here. On the one hand, adjectives agree in French, whereas they do not in English, which explains the erroneous plural on the adjective "complementary". On the other hand, "information" is uncountable in English, whereas the plural would be used in French (*des informations*).

This study has limitations, including the fact that it is based on a relatively small corpus[12] of 123 abstracts; therefore, some error types not included in our typology may have been overlooked. However, we are confident that similar investigations into larger corpora of abstracts would confirm our main findings, notably a convergence towards six predominant types of errors (with some degree of disciplinary variation). Although error dispersion and variability were seen amongst the 123 abstracts, determiners, syntax, choice of tenses (particularly a clear trend towards using the present perfect to describe methods), compound phrases, collocations/phraseology, and obscure formulations could be described as the main areas of concern.

Further, interestingly, a Geoscience abstract (http://www.theses.fr/2016LORR0342) did not have many errors (9 errors, word count = 272), but its style was consensually identified as "poor" by the four analysts, perhaps due to its excessively complex style (and possibly because of an extensive use of terminology). Therefore, it should also be granted that factors other than

[12] Previous research on the development of language proficiency tests based on data derived from corpora has used larger corpora. A notable example is Sharpling (2010), who used the British Academic Written English (BAWE) corpus to develop an English language test at the University of Warwick.

the errors identified in this article (i. e. text coherence and cohesion, thematic-rhematic structure, average length of sentences, terminological density, for example) may affect the clarity, readability and acceptability of a scientific abstract written in English.

Overall, very significant issues were noted in abstracts written by low-proficiency doctoral students, thereby limiting the potential impact of the research conducted by those doctoral students in their disciplinary communities. We tentatively attempted to define a metric of error frequency by dividing the total number of linguistic and stylistic errors by the word count of each abstract, thereby obtaining the *Comprehensive Error Ratio* or CER[13]. CERs significantly varied in our corpus, as can be seen from Table 9:

Table 9: Comprehensive error ratios (CERs) in the GC and the MEC

	GEOSCIENCE	MECH. ENGINEERING
Mean CER	5.79 %	7.89 %
Highest CER	15.09 %	18.29 %
Lowest CER	3.08 %	3.90 %
CER standard deviation	2.5	3.1

One can observe that mean CERs were relatively close in the GC and the MEC, which suggests that disciplinary variation in the overall number of errors was relatively limited between our two corpora. Although the range of CERs was extended in both corpora, one could reasonably assume that CERs above 6 %[14] reflected a very poor quality of language and style (22 abstracts out of 62 had CERs above 6 % in the GC, versus 42 out of 61 in the MEC, reflecting a higher concentration of errors in poorly written mechanical engineering abstracts). Having or helping doctoral students or junior researchers compute their CERs in first drafts of abstracts or articles while receiving scientific writing training may therefore help them become more aware of the need to mitigate their errors. Objectives in terms of CER reduction may be assigned to students over the course of a semester or an academic year. Combined with exposure to corpora reflecting the stylistic patterns generally accepted in the scientific community, notably through concordancing activities based on disciplinary corpora of research articles (Flowerdew 2015, Simard 2018) we believe that CER awareness-building could significantly improve the quality of scientific abstracts or articles. Results of studies on high-level doctorate students' writing (Charles 2011) and links to concordances from proficient student assignments such as those found in the BAWE corpus (Vincent/Nesi 2018) could also be used as alternative teaching tools.

However, "error is not sin": rather than being viewed as negative and penalizing, errors should be reconsidered as a positive contribution to language learning. We also need to keep in mind that the perception of errors varies according to one's institutional position in the ecosystem of research (Sionis 1993), and that native language researchers might not consider some errors in the same way as ESP/EAP instructors do. Another important remark by Hamilton (2015) is that grammar and lexis cannot constitute the only pedagogical objective when training learners of English, nor the only objective of error analysis. In this respect, we would concur with Vincent/Nesi (2018: para 28), when they point out: "Not all the problems we en-

[13] Relevant results for each abstract can be obtained from the authors upon request by email.

[14] As an example, the second Geoscience abstract had a CER of 8.02 %, with 21 errors for a total word count of 285.

counter [in student academic writing] can be classed as 'errors' in the sense of grammatical mistakes; some are simply unidiomatic or in an inappropriate register, often because of the writers' collocation choices." Further, grammar cannot be separated from text cohesion and coherence, also called "global errors" as they affect semantic links between sentences (Carter-Thomas 1998). Pragmatic considerations or text organization should also be taken into account for error analysis and categorization, which of course warrants further studies.

When considering a typology of errors in scientific writing, one should also accept the fact that, as suggested by Hynninen/Kuteeva (2017), perceptions of what constitutes "good" scientific writing may significantly vary between disciplines such as history and computer science, although a trend towards "a standard norm" based on understandability and clarity is also seen among various disciplinary discourse communities. McKinley/Rose (2018) thereby suggested that research journals should re-think their conceptualization of "error-free writing" in English for research purposes, notably by decoupling it from "nativeness". In this respect, developing typologies of errors may help non-native scientists mitigate the most frequently observed linguistic and stylistic issues. We hope that the tentative typology presented in this study may help them expand their proofreading skills through better awareness of frequent errors. Error typologies may therefore empower them to improve linguistic and stylistic choices, and ultimately become competent members of discourse communities in their disciplines.

References

Albert, Camille/Garnier, Marie/Rykner, Arnaud/Saint-Dizier, Patrick (2010): "Description et Annotation des Erreurs. Le Cas des Francophones s'Exprimant en Anglais." *Multilinguisme et Traitement Des Langues Naturelles*. Eds. Ismaïl Biskri/Adel Jebali. Québec: Presses de l'Université Du Québec. 54–70.

Biber, Douglas/Gray, Bethany (2010): "Challenging Stereotypes about Academic Writing: Complexity, Elaboration, Explicitness." *Journal of English for Academic Purposes* 9.1: 2–20.

Birch, Susan (1996): *Les Besoins Linguistiques des Scientifiques Français Publiant en Anglais: Analyse d'un Corpus de Premières Rédactions et de leur Correction*. Lille: ANRT, Université de Lille III.

Birch-Bécaas, Susan (1996): "French Researchers Publishing in English. An Analysis of a Corpus of First Drafts." *ASp. La Revue du GERAS* 11–14: 75–88. 02.03.2020 <https://journals.openedition.org/asp/3432>.

Bolton, Kingsley/Nelson, Gerald/Hung, Joseph (2002): "A Corpus-Based Study of Connectors in Student Writing: Research from the International Corpus of English in Hong Kong (ICE-HK)." *International Journal of Corpus Linguistics* 7.2: 165–182.

Bondi, Marina/Lorés Sanz, Rosa, eds. (2014): *Abstracts in Academic discourse*. Bern: Lang.

Buckingham, Louisa/Aktuğ-Ekinci, Duygu (2017): "Interpreting Coded Feedback on Writing: Turkish EFL Students' Approaches to Revision." *Journal of English for Academic Purposes* 26: 1–16.

Bychkovska, Tetyana/Lee, Joseph J. (2017): "At the Same Time: Lexical Bundles in L1 and L2 University Student Argumentative Writing." *Journal of English for Academic Purposes* 30: 38–52.

Carter-Thomas, Shirley (1998): *Conference at the Université de Bretagne Occidentale (7ème Journée ERLA/GLAT) "Erreurs Locales et Erreurs Globales: Une Contribution à l'Analyse Textuelle de l'Anglais Scientifique", Rennes, 4 June 1998*. 15.03.2019 <https://www.researchgate.net/publication/32221355_Erreurs_locales_et_erreurs_globales_une_contribution_a_l%27analyse_textuelle_de_l%27anglais_scientifique>.

Charles, Maggie (2011): *Corpus Linguistic Conference at the University of Birmingham "Making Concessions in Academic Writing: A Corpus Study of Patterns and Semantic Sequences", Birmingham, 20.–22. July 2011*. 02.03.2020 <https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/paper-88.pdf>.

Corder, Stephen Pit (1967): "The Significance of Learners' Errors." *International Review of Applied Linguistics in Language Teaching* 5.4: 161–170.

Corder, Stephen Pit (1973): *Introducing Applied Linguistics*. London: Penguin.

Cortes, Viviana (2013): "The Purpose of This Study Is To: Connecting Lexical Bundles and Moves in Research Article Introductions." *Journal of English for Academic Purposes* 12.1: 33–43.

Doushaq, Hufeeq H. (1986): "An Investigation into Stylistic Errors of Arab Students Learning English for Academic Purposes." *English for Specific Purposes* 5.1: 27–39.

Ferris, Dana (2011): *Treatment of Error in Second Language Student Writing*. Ann Arbor: University of Michigan Press.

Ferris, Dana/Roberts, Barrie (2001): "Error Feedback in L2 Writing Classes: How Explicit Does It Need to Be?" *Journal of Second Language Writing* 10.3: 161–184.

Flowerdew, Lynne (2000): "Investigating Referential and Pragmatic Errors in a Learner Corpus." *Rethinking Language Pedagogy from a Corpus Perspective*. Eds. Lou Burnard/Tony McEnery. Frankfurt a. M.: Lang. 145–154.

Flowerdew, Lynne (2015): "Using Corpus-Based Research and Online Academic Corpora to Inform Writing of the Discussion Section of a Thesis." *Journal of English for Academic Purposes* 20: 58–68.

Gaillat, Thomas (2013): "Annotation Automatique d'un Corpus d'Apprenants d'Anglais avec un Jeu d'Étiquettes Modifié du Penn Treebank". *Proceedings of TALN 2013. Les Sables D'Olonne, 17.–21. June 2013:* 271–284. 02.03.2020 <https://www.aclweb.org/anthology/F13-1020>.

Gilquin, Gaëtanelle/Granger, Sylviane/Paquot, Magalie (2007): "Learner Corpora: The Missing Link in EAP Pedagogy." *Journal of English for Academic Purposes* 6.4: 319–335.

Gledhill, Christopher J. (2000a): "The Discourse Function of Collocation in Research Article Introductions." *English for Specific Purposes* 19.2: 115–135.

Gledhill, Christopher J. (2000b): *Collocations in Science Writing*. Tübingen: Narr.

Granger, Sylviane/Tyson, Stephanie (1996): "Connector Usage in the English Essay Writing of Native and Non-Native EFL Speakers of English." *World Englishes* 15.1: 17–27.

Halliday, M. A. K./Martin, J. R. (1993): *Writing Science: Literacy and Discursive Power*. London: Routledge.

Hamilton, Clive E. (2015): *Mapping English L2 Errors: An Integrated System and Textual Approach*. Doctoral Dissertation, Université Sorbonne Nouvelle – Paris 3. <https://hal.archives-ouvertes.fr/tel-01378302>, accessed on April 2018.

Hinkel, Eli (2004): "Tense, Aspect and the Passive Voice in L1 and L2 Academic Texts." *Language Teaching Research* 8.1: 5–29.

Hyland, Ken (2008): "As Can Be Seen: Lexical Bundles and Disciplinary Variation." *English for Specific Purposes* 27.1: 4–21.

Hynninen, Niina/Kuteeva, Maria (2017): "'Good' and 'Acceptable' English in L2 Research Writing: Ideals and Realities in History and Computer Science." *Journal of English for Academic Purposes* 30: 53–65.

James, Carl (1998): *Errors in Language Learning and Use: Exploring Error Analysis*. London: Routledge.

Kübler, Natalie/Mestivier, Alexandra/Pecman, Mojca/Zimina, Maria (2016): "Exploitation Quantitative de Corpus de Traductions Annotés Selon la Typologie d'Erreurs pour Améliorer les Méthodes d'Enseignement de la Traduction Spécialisée." *Actes des 13èmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2016)*: 7–10.

Luzón, María José (2009): "The Use of We in a Learner Corpus of Reports Written by EFL Engineering Students." *Journal of English for Academic Purposes* 8.3: 192–206.

McKinley, Jim/Rose, Heath (2018): "Conceptualizations of Language Errors, Standards, Norms and Nativeness in English for Research Publication Purposes: An Analysis of Journal Submission Guidelines." *Journal of Second Language Writing* 42: 1–11.

Mignot, Elise (2001): *Recherches sur les Noms Composés de Type Nom + Nom en Anglais Contemporain*. PhD thesis: Paris 4 <http://www.theses.fr/2001PA040192>, accessed on April 2018.

Norrish, John (1983): *Language Learners and Their Errors*. London: Macmillan.

Palmer, Harold E. (1917): *The Scientific Study and Teaching of Languages*. Oxford: Oxford University Press.

Palmer, Harold E. (1921): *The Principles of Language Study*. New York: World Book Company.

Palmer, Harold E. (1924): *Memorandum on Problems of English Teaching in the Light of a New Theory*. Tokyo: Institute for Research in English Teaching.

Pan, Fan/Reppen, Randi/Biber, Douglas (2016): "Comparing Patterns of L1 Versus L2 English Academic Professionals: Lexical Bundles in Telecommunications Research Journals." *Journal of English for Academic Purposes* 21: 60–71.

Payre-Ficout, Coralie/Chevrot, Jean-Pierre (2004): "La Forme contre l'Usage. Étude Exploratoire de l'Acquisition du Prétérit Anglais par des Apprenants Français." *Lidil. Revue de Linguistique et de Didactique Des Langues* 30: 101–115.

Rowley-Jolivet, Elizabeth (2001): "Activating the Passive – A Comparative Study of the Passive in Scientific Conference Presentations and Research Articles." *Recherche et Pratiques Pédagogiques en Langues de Spécialité. Cahiers de l'Apliut* 20.4: 38–52.

Sampson, Andrew (2012): "Coded and Uncoded Error Feedback: Effects on Error Frequencies in Adult Colombian EFL Learners' Writing." *System* 40.4: 494–504.

Sharpling, Gerard Paul (2010): "When BAWE meets WELT: The Use of a Corpus of Student Writing to Develop Items for a Proficiency Test in Grammar and English Usage." *Journal of Writing Research* 2.2: 179–195.

Simard, Laura-May (2018): "Le Corpus comme Aide à la Rédaction de Résumés Scientifiques pour des Étudiants LANSAD: Une Approche Comparative." *ASp La Revue du GERAS* 73: 75–104. 26.04.2019 <https://journals.openedition.org/asp/5122>.

Sionis, Claude (1993): "Relativité de l'Erreur en Anglais de Spécialité des Sciences et des Techniques." *Cahiers de l'APLIUT* 13.1: 95–108.

Sionis, Claude (1995): "Communication Strategies in the Writing of Scientific Research Articles by Non-Native Users of English." *English for Specific Purposes* 14.2: 99–113.

Storch, Neomy/Wigglesworth, Gillian (2010): "Learners' Processing, Uptake and Retention of Corrective Feedback on Writing: Case Studies." *Studies in Second Language Acquisition* 32.2: 303–334.

Swales, John M./Feak, Christine B. (2009): *Abstracts and the Writing of Abstracts*. Ann Arbor: University of Michigan Press.

Truscott, John (2007): "The Effect of Error Correction on Learners' Ability to Write Accurately." *Journal of Second Language Writing* 16.4: 255–272.

Truscott, John/Yi-ping Hsu, Angela (2008): "Error Correction, Revision, and Learning." *Journal of Second Language Writing* 17.4: 292–305.

Van Beuningen, Catherine G./De Jong, Nivja H./Kuiken, Folkert (2012): "Evidence on the Effectiveness of Comprehensive Error Correction in Second Language Writing." *Language Learning* 62.1: 1–41.

Vincent, Benet/Nesi, Hilary (2018): "The BAWE Quicklinks Project: A New DDL Resource for University Students" *Lidil. Revue de linguistique et de didactique des langues* 58. 02.03.2020 <http://journals.openedition.org/lidil/5306>.

Wilkinson, A. M. (1992): "Jargon and the Passive Voice: Prescriptions and Proscriptions for Scientific Writing." *Journal of Technical Writing and Communication* 22.3: 319–325.

Appendix 1: Corpora of abstracts from www.theses.fr

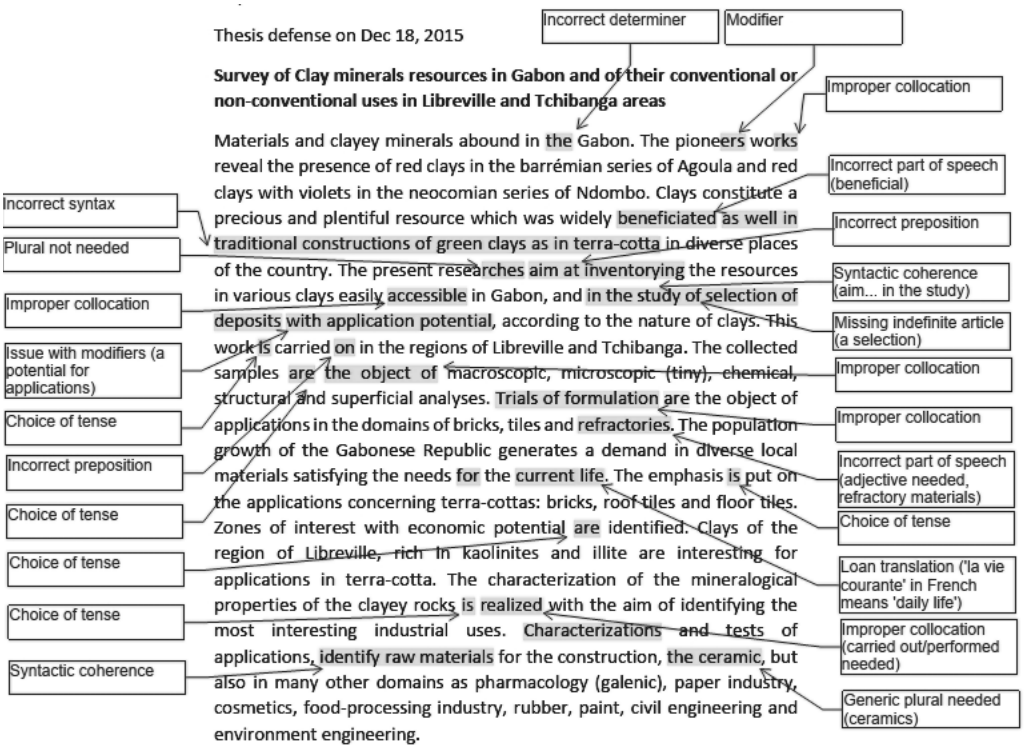
Abstract references are arranged by date of publication. To obtain the full URL of each abstract, add http://www.theses.fr/ before each code.

Geoscience	**Corpus**	**(GC)**
2004INPL032N	2012NICE4048	2016LORR0193
2006AIX30068	2013STRAH023	2016LORR0206
2007NAN10018	2013LORR0361	2016LORR0270
2008NAN10075	2013STRAH006	2016LORR0342
2008STR1GE04	2013PA077153	2016GREAU025
2008MON20230	2014MON20057	2017LORR0153
2009GRE10104	2014STRAH019	2017LORR0264
2009STRA6056	2014MON20105	2017PA066220
2009PA077034	2014MON20149	2017PA066571
2009NICE4085	2014MON20133	2017PA066476
2010PA077188	2014STRAH008	2017PA066289
2010LIL10068	2015MONTS036	2017LORR0360
2011NAN10149	2015LORR0309	2017LORR0059
2011PA077037	2016LORR0026	2017LORR0056
2012PA066494	2016LORR0058	2017LORR0280
2012PA077078	2016LORR0063	2017LORR0383
2012LORR0292	2016LORR0097	2017LORR0213
2012STRAH005	2016LORR0134	2017LORR0334
2012LORR0142	2016PA066010	2017LORR0118
2012NICE4022	2016LORR0019	
2012NICE4109	2016LORR0148	

Mechanical Engineering	**Corpus**	**(MEC)**
2003VALE0003	2007INPT008H	2015EMAC0001
2003ISAT0002	2008REIMS014	2015ESAE0009
2003VALE0043	2008INPT051H	2015ENAM0012
2003VALE0014	2008ISAT0039	2015EMAC0008
2003ECAP0916	2008TOU30253	2015ISAL0010
2004VALE0009	2008GRE10016	2015SACLC020
2004VALE0030	2009DENS0026	2015ISAL0105
2004ISAL0005	2009ISAL0011	2016ENAM0001
2004ISAL0090	2010ECDN0030	2016ENAM0073
2005DENS0041	2010ENAM0032	2016ENAM0033

2005VALE0011	2011ECDN0043	2016LYSEI141
2005TOU30256	2011ECDN0050	2017ISAR0004
2005ISAT0033	2011TOU30050	2017BRES0104
2006CHAMS001	2011ENAM0020	2017SACLN059
2006BESA2033	2011STRA6017	2017EMAC0003
2006ENAM0029	2011ECDL0045	2017LYSEC032
2006VALE0007	2012ENAM0022	2017SACLC04
2006NANT2080	2013ISAL0122	2017NORMIR05
2007DENS0038	2013ORLE2065	2017VALE0012
2007ISAT0012	2014ESAE0047	
2007TOU30023	2014EMAC0014	

Appendix 2: Example of an annotated Geoscience abstract



Dr. Anthony Saber
Audrey Cartron
Claire Kloppmann-Lambert
Céline Louis
École normale supérieure (ENS) Paris-Saclay, France,
Département des langues
Department of languages, DASP research unit
4 avenue des sciences
91190, Gif sur-Yvette
France
anthony.saber@ens-paris-saclay.fr