Lingue e Linguaggi
Lingue Linguaggi 29 (2019), 557-579
ISSN 2239-0367, e-ISSN 2239-0359
DOI 10.1285/i22390359v29p557
http://siba-ese.unisalento.it, © 2019 Università del Salento

CREATING LOW-TECH HIGH-RELEVANCE PERSONALIZED LANGUAGE CORPORA WITH SCIENCE POSTGRADUATES

CARMELA MARY WHITE
UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Abstract – This paper proposes a technically simple way for students to exploit individual linguistically specialized corpora which are directly relevant to their professional learning curve. Established science researchers today use highly specialized forms of professional ESP communication in spoken, written, or mixed media, although the central vehicle for research validation remains the research article. Since academic success is measured in publications, postgraduates aspiring to a scientific career, while learning to design and perform research activities, must also gain communicative proficiency. Raising students' awareness of the linguistic characteristics and complex unwritten rules of research article construction can boost their communicative performance. Language acquisition theorists advocate the use of specialist corpora for acquiring specialist discourse competence, a theory supported by experimental studies involving the use and/or creation of personalized language corpora by students using concordancing software. However, science postgraduates can be reluctant to spend time learning to use even a simple concordance tool. Following a brief theoretical introduction and literature review, the author describes a teaching experience with a physics PhD group in which students created and studied personalized corpora of specialist research articles without a concordancing tool, whilst the teacher used concordancing to create tailor-made classroom activities based on the collective class corpus. Students prepared for peer-to-peer communication of research by analysing linguistic features, practising problematic points and comparing the rhetorical structure of their corpus articles with a generic model, thus building up a complex picture of how experts in their field do not simply convey information but work to persuade and reveal their stance.

Keywords: ESP; postgraduate EST course; personalized specialist field corpus; classroom activities.

1. Introduction

The overall aims of this paper are to explain why postgraduate science students need specialist language corpora and to identify and propose a technically simple method for getting students to create and use individual linguistically

specialized corpora as an integral part of specialist English language courses. University administrators looking to improve efficiency and reduce costs ask why they should employ resources specifically for English for Science and Technology (EST) courses, rather than aim for general linguistic academic competency, to be taught in courses shared by all disciplines (Hyland 2009). To provide an insight as to *why* science students need purposely designed specialist EST courses, this introductory section first draws from the theoretical background in a variety of fields to explore both the nature of the target communicative competence and the relevance of specialist corpora to EST courses. This is followed by a short review of some recent papers to see *how* specialized text corpora have been successfully employed in ESP/EST teaching contexts, in doing so, a common obstacle reported by various authors is identified. The author's aim in the remaining sections of the paper is to suggest a possible solution that might encourage more EST practitioners to experiment with personalized corpora.

1.1. The need for purposely-designed specialist EST courses

Established hard science researchers use highly specialized forms of professional communication in spoken, written, or mixed media. Participants of a survey (White, forthcoming) carried out amongst teaching staff at the former Faculty of Science at Bari University report an increasing use of spoken English in their work and a variety of written genres, although the central vehicle for research validation remains the formal research article (RA) with more than 90% of survey participants declaring they publish exclusively in English. The academic pressure on individuals and institutions to perform well has never been greater, success being measured predominantly in terms of international publications (see, for example, the websites of the Italian ANVUR and equivalent UK REF, responsible for higher education institute research evaluation). Clearly, postgraduates aspiring to a career in science, while learning to design and perform research activities, must also gain communicative proficiency.

Hyland (2009) emphasizes the need for students to learn specialist discourses, but notes a tendency on the part of tutors and administrators to misrepresent academic literacy as a "single, overarching literacy which students have failed to master before they get to university", a shortcoming seen as easily remedied by "a few top-up English classes". This ideology, he warns,

> transforms literacy from a key area of academic practice, how we construct ourselves as credible linguists, psychologists or whatever, into a kind of add-on to the more serious activities of university life. (Hyland 2009, p. 9)

Hyland here touches on the important sociolinguistic aspects of academic literacy: unless students are given an opportunity to conceptualize the specific epistemological frameworks of academic research, they will have difficulty acquiring the necessary communication skills to grow professionally and enter their target community of practice. In community of practice terms (Wenger *et al*. 2002), postgraduate students are novices beginning initiation into a relatively exclusive community of practice, where knowledge of the community discourse plays a key role in community gatekeeping (Hyland 2009, p. 5). Thus, knowledge of specialist community discourse is an essential part of a postgraduate's target communicative competence.

1.2. Target communicative competence

EST teachers need to familiarize themselves with the cognitive frames of science to understand how the message to be conveyed and the chosen form of expression are linked. Tarantino recommends that applied linguists should, 'immerse themselves into the culture of the disciplinary community and become conversant with its languages and conventions' (Tarantino 2011, p. 180), since its language is inextricably bound to the conceptual framework both of scientific investigation in general and that of the specific discipline. She reminds us that neither everyday language nor even scientific outreach prose can be equated to scientific discourse because

> the content matter of specialist reports builds on different sources of knowledge, hence, it cannot be properly understood by people not trained in the specific field of research and application. Lay people can repeat technical terms in speech or writing. However, since they are unaware of the non-verbal dimensions which technical expressions embody, they lack the knowledge required to evaluate the reliability of a scientific text, criticise or expand its content. (Tarantino 2004, p. 84)

The complex scientific cognitive frames which science students must acquire have been developed over centuries, but for the purposes of this paper can be summarized in the following short tenets.

Firstly, scientists formulate and test measurable hypotheses about the world surrounding us. Early in their university careers, students learn how to produce qualitative and quantitative empirical data (mathematical studies, laboratory experiments) and begin the process of learning to write science (laboratory reports).

Secondly, the data thus produced are presented to support statements aimed at contributing to collectively accepted knowledge. The rationale behind the laboratory report format and content becomes clearer to students as they gradually acquire a scientific mindset, aided by a shift in reading from text books to research papers during their postgraduate studies.

Finally, every professional scientist acknowledges that the collectively accepted knowledge of his/her field, is always open to modification, through peer evaluation and further hypothesizing and testing, in direct contrast with the layman's tendency to view scientific knowledge as written in stone. These general scientific tenets model the language of the scientific discourse community as a whole, but "gaining specialized knowledge of a disciplinary community's genres and mastering them presupposes taking on the discipline's identity" (Dressen-Hammouda 2008, p. 234), so that students need to be exposed to specialist genres within their own discipline.

Numerous aspects of scientific discourse are described in the literature, including its grammatico-syntactical and lexical (Halliday 2004), rhetorical functional (Widdowson 1979, Swales, 1990) and metadiscoursal (Hyland, Tse 2004) characteristics and even the rationale behind its historical development (Gross 2002), which together set it apart from other forms of English. Halliday (2004, p. 162) identifies seven problematic grammatico-syntactical areas which are stumbling blocks for the uninitiated.

One of the most evident features is EST's tendency towards nominalization, which Halliday traces to the logical linear Theme-Rheme construction of arguments in research texts, consisting typically of a long sequence of connected steps. At any point the author may require a large number of previous concepts to support his/her next argument, therefore "the only way to package a piece of argument so that it becomes a natural Theme of a clause is to turn it into a nominal group" (Halliday 2004, p. 125).

The need to express complex new concepts thus leads to another feature of EST text: high lexical density. The following is a typical example characterized by a high lexical density and nominalization, coupled with a relatively simple clause structure.

> Among the most robust and sensitive trace-gas detection techniques, quartz-enhanced photo-acoustic spectroscopy (QEPAS) is capable of record sensitivities with a compact and relatively low-cost acoustic detection module (ADM). (Spagnolo *et al.* 2013, p. 1)

Another of the foremost characteristics of an EST text is its multimodality. Driven by the need to summarize bulk data efficiently and meaningfully, a considerable quantity of information is given through visuals, which have their own conventions and "grammar" (Kres, van Leewen 1996).

Furthermore, cutting edge science articles have long attracted the interest of linguistics for their lexical creativity (see Halliday 2004, Chapter 2) since new concepts require language to express them: the title of the paper cited above is a good example: "THz quantum cascade laser-based quartz enhanced photo-acoustic Sensor", which requires considerable "unpacking" or "decoding". The uninitiated would need each decoded element to be explained

with at least one sentence to even begin to understand all the information conveyed to the expert reader.

The rhetorical conventions developed for each section of the standard Introduction Methods Results and Discussion (IMRAD) RA format form another layer of unwritten code. The various sections "perform different rhetorical functions and thus require different linguistic resources to realize those functions" (Swales 1990, p. 136), such as the need to juggle expertly past, present and perfect tenses in the Create A Research Space (CARS) model introduction section, the extensive use of the past simple passive in the method section, the use of epistemic modality in discussing results.

Metadiscoursal features are also frequent in EST texts, whether aimed at negotiating collective consensus, including hedging and other mechanisms to indicate stance, or helping the reader navigate the text. These structures, too, are "intimately linked to the norms and cultural expectations of particular cultural and professional communities" (Hyland, Tse 2004, p. 175). Thus, if our aspiring scientists are to succeed, they need to be aware of the conceptual frames, rhetorical structure and appropriate metadiscoursal features through which to share their work and gain acceptance within their specific target discourse community.

1.3. The case for using individual corpora in EST courses

We have seen how various studies suggest that raising students' awareness of the complexities of the unwritten rules can boost their communicative performance. Does work on corpora provide a suitable means of achieving this end?

An argument strongly in favour of using corpora directly with students is that corpora consist of authentic materials. Various language acquisition theories, including lexical bundles (Hyland 2008) and lexical priming (Hoey 2005), suggest that the use of authentic material in corpora provides excellent material for learners. Moreover, language theorists have further advocated the use of specialist corpora for acquiring specialist discourse competence. Indeed, specialist texts related to the student's own field of research surely fulfil Krashen's comprehensible input requirements (Krashen 1987) and those of the Vygotskian Zone of Proximal development (ZPD) theory (Lantolf, Poehner 2014). For a more detailed review of corpus-based language research and its application to language pedagogy see McEnery and Xiao (2010).

The literature provides a number of interesting experimental studies in ESP teaching contexts, using concordancing programmes with students and reporting positive outcomes in the main.

Lee and Swales report on a corpus-based EAP course for NNS doctoral students which they describe as an exercise in "technology-enhanced rhetorical

consciousness-raising" (Lee, Swales 2006, p. 58). In this course, students compared their own writing with the expert texts. The authors discuss how corpora can be used in the classroom, concluding that the degree of specialization of the corpus used should depend on the level of disciplinary acculturation of the course participants. They also note that concordancing works better at lexico-grammatical and phraseological rather than structural levels (such as audience analysis and paper structure). This suggests that activities exploiting other methods of analysis should be developed alongside concordancing activities in an EST class, since the target communicative competence for PhD science students requires detailed knowledge of specialist RA structure.

Charles (2012) investigates the feasibility and value of an approach to teaching postgraduate EAP writing in which students construct and examine DIY discipline-specific corpora. She argues that the possible benefits outweigh the disadvantages which must characterize a comparatively diminutive corpus:

> First, it is unlikely that even a large general corpus will provide adequate data to respond to the highly discipline-specific queries of specialist students. Thus users may find that there are few or no examples, or that the examples retrieved are irrelevant, or even misleading. Second, I would suggest that the process of building their own corpus allows students to achieve deeper and more critical insights into the nature of corpus data itself. This understanding helps them to interpret corpus data more perceptively and to gain a greater appreciation of the pitfalls as well as the benefits of the approach. (Charles 2012, p. 94)

Milizia (2013), instead, introduced political science undergraduates to concordancing software using a specialist corpus. She observes that these students are interested in analysing political speeches, enthusiastic about working with authentic materials, and like to "get their hands on corpora and concordances themselves and find out about language patterning" (Milizia 2013, p. 158). In fact, students who completed the courses are reported in all three studies to have been enthusiastic.

However, unlike Milizia's undergraduates, who were both predisposed to language analysis and excited by their first real approach to research, only a handful of the postgraduates completed Lee and Swales' course, and although Charles finds most students were able to compile their own "quick and dirty" corpus, she reports that less than a third followed all the classes. Comments collected by the authors suggest that an important factor in the drop-out rate could have been the time factor: some students may have found the pay-off too high in terms of time/perceived benefit. Indeed, it is the experience of this author, when asking previous cohorts of science postgraduates in an anonymous course feedback questionnaire whether they would be interested in

in learning to use a concordance programme, that few expressed a willingness to dedicate time to this activity.

This is corroborated by Adams' observations (2006), the purpose of whose research was to trial the concordancing approach in both discipline specific and general academic contexts. Adam remarks that hard sciences PhD students, who are already subject to time pressure, may be unwilling to dedicate time to learning to use a concordancing tool, adding that

> Software in this field is not usually designed for language learners to use: it is typically designed for language researchers and is often complex to operate. It also uses jargon familiar only to linguists and language specialists, adding to the time and effort necessary for students and teachers to learn. (Adams 2006, p. 6)

This points to an unresolved question for teachers of EST: might it be possible to take advantage of the benefits offered by individual specialist corpora in an EST context without the need for students to learn to use a concordancing tool?

In the light of the theoretical and pedagogical background of EST and the use of corpora discussed above, the author describes below a teaching experience aimed at exploring an alternative way to get individual students to create for themselves a technically simple, but linguistically specialized, language corpus. In Section 2 the author describes the constraints, overall aims and choice of materials for the course, while Section 3 gives details of the set-up of the individual and class corpora. Section 4 provides examples of teaching activities developed using the RA corpora, followed by brief reflections on the experience and possible implications for EST course development.

2. Course constraints and choice of materials

The activities which will be described in Section 3 were developed for a brief mandatory (16 hour in-class) language course offered each year by the physics department at Bari University to first year PhD students. The students who follow this course are generally highly motivated, though very pressed for time.

The data used in this paper refer specifically to the 32nd doctoral cycle course. Two of the 13 participants of the 32nd cycle were not native Italian speakers; in a brief written and spoken entry test one participant performed at the top of the CEFR level B1, eight in the B2 level range, and the remaining four at C1 level or over. A needs analysis revealed a wide variety and high level of complexity of the communicative tasks which these physics PhD students may be expected to perform in English during their PhD studies. [1]

[1] Including reading specialist RAs, lab manuals and emails, writing applications for stages abroad, preparing periodic research progress reports, or even collaboration writing RAs. Oral and/or mixed media tasks may include Skype interviews, meetings, poster preparation and conference poster

However, the time constraint necessarily confined the scope of the course to only a limited part of these tasks.

Subsequently, participants' own research was selected as the main focus, it was nonetheless agreed, together with the head of the Doctoral school, that the participants needed all round linguistic support. The following teaching strategies were therefore selected to work towards the overall goal of improving students' communicative performance in oral presentations supported with slides:

- class activities promoting fluency and practising delivery techniques and pronunciation;
- vocabulary expansion of specialist terminology and its collocations;
- vocabulary expansion of metadiscoursal chunks of both audience/reader – presenter/writer relations and speaker/writer's stance;
- exposure to authentic materials in the form of specialist RAs and audiovisual;
- material available on internet; the recreation (as far as possible) of conference conditions for the final presentations, with audience tasks and debriefing.

To provide authentic oral texts for the whole class, the author selected a number of audio and video materials freely available on the internet, ranging from undergraduate video lecture excerpts and podcasts to Doumont's (2010) excellent scientific presentations unit on the *Nature Scitable* website.[2] Such materials are very useful because they provide examples of generic scientific content and/or oral research presentations, yet they cannot provide examples of specialist discourse for every field. For the specific discourse of their individual fields, authentic specialist field materials therefore took the form of RAs whose content was directly relevant to the individual, thus providing the best available alternative to specialist oral text.

The reader may require a further explanation regarding the suitability of choosing *written* articles as material for an *oral* presentation course (beyond their highly relevant specialist discourse and content). Firstly, it should be remembered that an oral presentation is a multimodal form of communication, where the more formal register of written RAs, duly compacted, has its place in the slides. The course participants all had to produce slides, which of course

sessions, conference contributions, participation and notetaking in seminars and international summer schools, daily ELF communication with colleagues at home or in stages abroad.

[2] The complete ebook is available at https://www.nature.com/scitable/ebooks/english-communication-for-scientists-14053993, while the 3 videos can be viewed at: https://www.nature.com/scitable/content/john-s-video-14024244, https://www.nature.com/scitable/content/marie-s-video-14033591, https://www.nature.com/scitable/content/jean-luc-s-video-14031500

requires the ability to produce suitable written sentences and subsequently to reduce them to bullet point format as needed. Secondly, as Hyland (2009) points out, conference presentations are more tightly organized and patterned than conversational speech. Indeed, differences between academic writing and speech should be seen on a continuum, rather than as clear-cut dichotomies. Thirdly, another important feature is shared by RAs and oral presentations: the interaction of text (whether spoken and/or written) and visuals. Hyland sums this concept up as follows:

> the complicating role of multimodal semiotics in spoken presentations of various kinds further undermin[es] a direct spoken-written split in communicative features. The non-verbal dimension of academic speech (and writing) in both constructing and conveying information of various kinds is substantial…. (Hyland 2009, p. 24)

Finally, the general RA format is followed more or less strictly in presentations (Mariotti 2012, p.67) and thus, the RA corpora provided important material for comparison, allowing similarities and differences in purpose, audience, structure, language and modality between the dominant RA, and the poster and presentation genres to be discussed in class.

A week before the course started, the 14 participants were asked to find two or more RAs in digital format, thus creating the basis of a personal corpus of specialist text. The criteria given to guide their choice were that the RAs should be of direct relevance to their own studies and written by acknowledged experts in their specialist field. Whether or not they had already studied them for the content was irrelevant.

The texts were vetted for suitability by the teacher. In some cases, students had selected reviews rather research articles, in which case they were asked to add another text presenting original research. A few texts had English language issues, and again the students were asked to add another text, preferably with at least one NS author, and warned that while the validity of the content and organization of the text were not in question and could therefore be evaluated, it could not be considered a reliable language model.

The resulting 28 articles formed the small class corpus of specialized text. Each student used both a digital and paper copy of their own texts. The purpose of this article is not to describe in detail the whole course or all the materials and activities, but to focus on the activities performed with the individual embryonic RA corpora and the class corpus, which are discussed in the next section.

3. Individual RA corpora and related class activities

The activities carried out using the RAs can be divided into individual, group or class exercises, or according to whether they are aimed at rhetorical-function- structural or grammatical-lexical analysis and practice. As stated above, the goal was to allow students to use their RAs as a corpus for rudimentary analysis without having to download and learn to use a concordancing tool. They either used built-in functions of standard pdf readers to analyse their individual corpora or performed exercises manually or orally.

3.1. Phase 1. Getting started with the individual RA corpora

Private study. As soon as their RAs had been vetted, students were instructed to familiarize themselves with their articles by:

- skim reading the abstract and conclusion and examining any visuals;
- reading the whole article, once without, then a second time with a dictionary, listing vocabulary which was new to them;
- using the pdf reader's 'Find' tool to check for more examples of the new vocabulary in their own texts.

Classroom activities. In pairs or small groups, students were asked to summarize briefly one of their chosen RA orally and explain in what way it was relevant to their own studies. They were then asked to discuss the following questions:

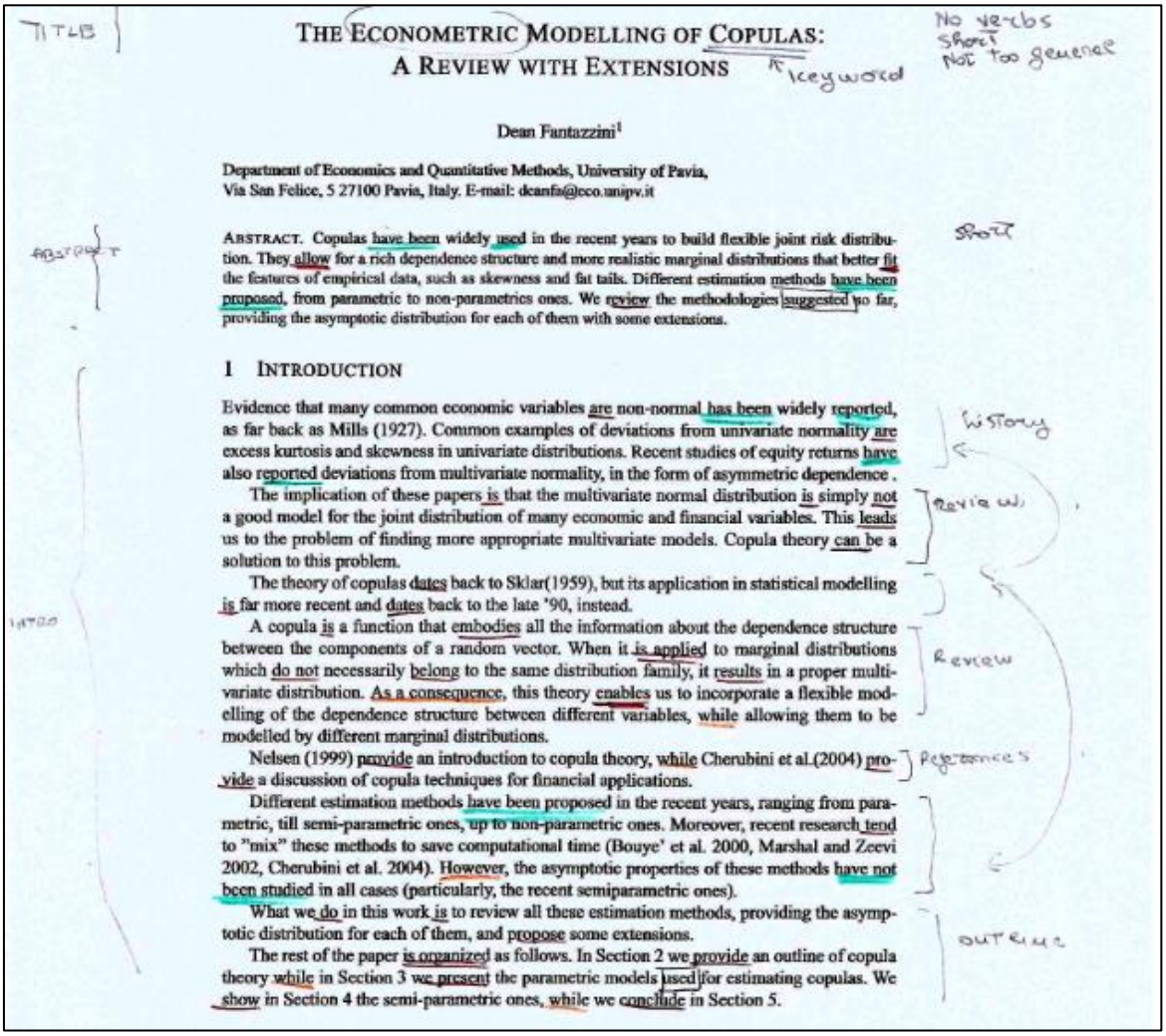
- Why do scientists need to write and present their research?
- Why do scientists need to write and present their research in English?
- What similarities and differences can you think of between a research article and a presentation? Consider content medium audience language organization.

This was followed by a class feedback session. A generalized model of the organization in sections of an RA was presented, which the students then quickly compared with their chosen texts, followed by further feedback.

3.2. Phase 2. Tense analysis of individual RA corpora

Private study. Students were first shown in class how to perform the tense analysis. They had the choice of doing this either manually using the paper copy, or digitally using their pdf reader (only possible with paid features of the software) or, possibly, by first laboriously moving the text into a Word file. Examples of both types were provided (Figures 1 and 2).

The analysis consisted in underlining each tense in a different colour, single underlining for active and double underlining for passive, to be completed for the abstract, introduction, and at least the first and final pages of every other section. Although this type of exercise may take some time, the coloured underlining makes the pattern of tense usage stand out clearly on the page, which is otherwise difficult to visualize. During the tense analysis, students were also encouraged to identify and mark reduced relative clauses, a high-frequency structure used in scientific prose to condense text.



THE ECONOMETRIC MODELLING OF COPULAS:
A REVIEW WITH EXTENSIONS

Dean Fantazzini[1]

Department of Economics and Quantitative Methods, University of Pavia,
Via San Felice, 5 27100 Pavia, Italy. E-mail: deanfa@eco.unipv.it

ABSTRACT. Copulas have been widely used in the recent years to build flexible joint risk distribution. They allow for a rich dependence structure and more realistic marginal distributions that better fit the features of empirical data, such as skewness and fat tails. Different estimation methods have been proposed, from parametric to non-parametrics ones. We review the methodologies suggested so far, providing the asymptotic distribution for each of them with some extensions.

1 INTRODUCTION

Evidence that many common economic variables are non-normal has been widely reported, as far back as Mills (1927). Common examples of deviations from univariate normality are excess kurtosis and skewness in univariate distributions. Recent studies of equity returns have also reported deviations from multivariate normality, in the form of asymmetric dependence .

The implication of these papers is that the multivariate normal distribution is simply not a good model for the joint distribution of many economic and financial variables. This leads us to the problem of finding more appropriate multivariate models. Copula theory can be a solution to this problem.

The theory of copulas dates back to Sklar(1959), but its application in statistical modelling is far more recent and dates back to the late '90, instead.

A copula is a function that embodies all the information about the dependence structure between the components of a random vector. When it is applied to marginal distributions which do not necessarily belong to the same distribution family, it results in a proper multivariate distribution. As a consequence, this theory enables us to incorporate a flexible modelling of the dependence structure between different variables, while allowing them to be modelled by different marginal distributions.

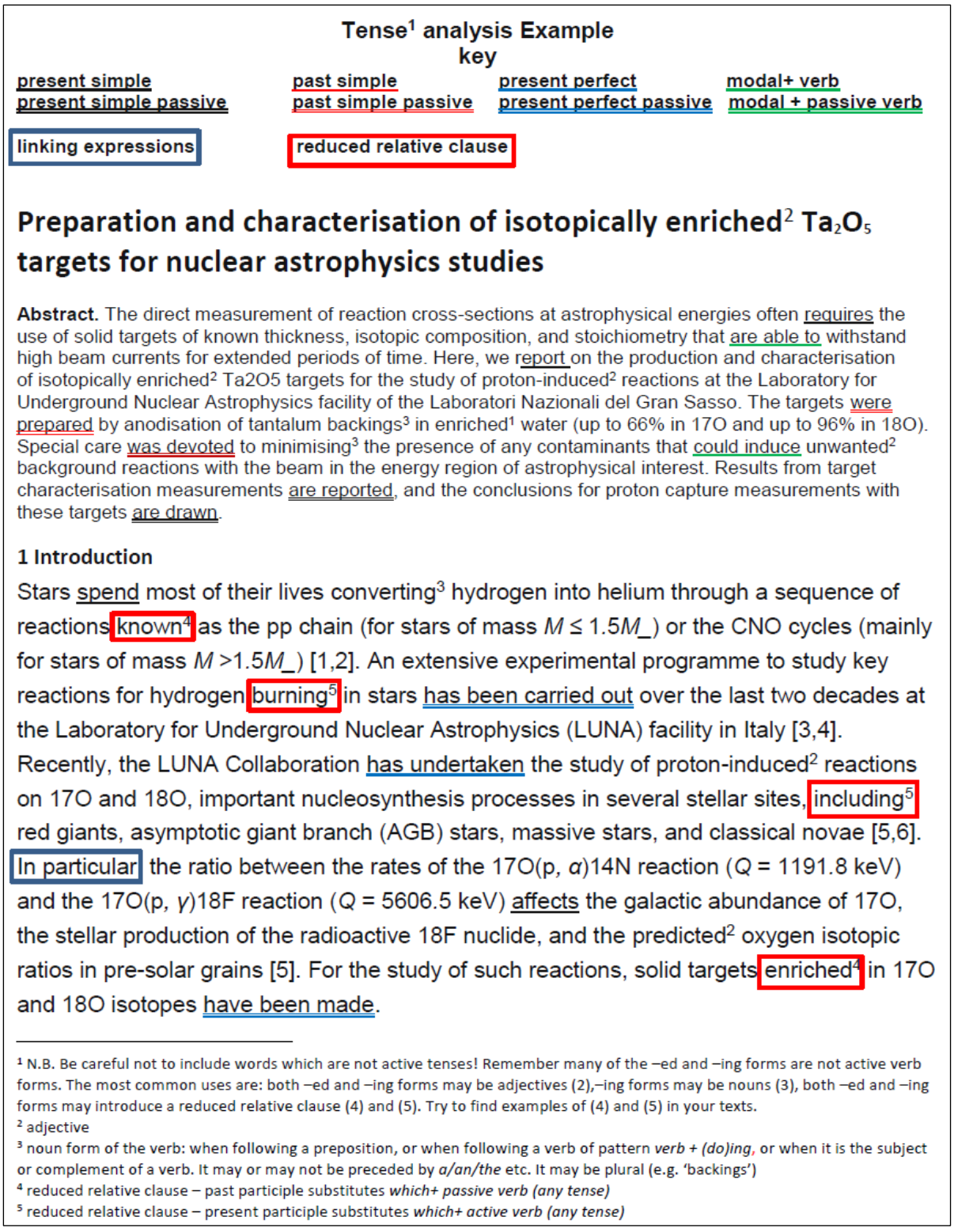
Nelsen (1999) provide an introduction to copula theory, while Cherubini et al.(2004) provide a discussion of copula techniques for financial applications.

Different estimation methods have been proposed in the recent years, ranging from parametric, till semi-parametric ones, up to non-parametric ones. Moreover, recent research tend to "mix" these methods to save computational time (Bouye' et al. 2000, Marshal and Zeevi 2002, Cherubini et al. 2004). However, the asymptotic properties of these methods have not been studied in all cases (particularly, the recent semiparametric ones).

What we do in this work is to review all these estimation methods, providing the asymptotic distribution for each of them, and propose some extensions.

The rest of the paper is organized as follows. In Section 2 we provide an outline of copula theory while in Section 3 we present the parametric models used for estimating copulas. We show in Section 4 the semi-parametric ones, while we conclude in Section 5.

Figure 1
Screen shot of a page of RA analysis with manual annotation produced by a student, provided to course participants as an example to follow.



Tense[1] analysis Example
key

present simple
present simple passive
past simple
past simple passive
present perfect
present perfect passive
modal+ verb
modal + passive verb
linking expressions
reduced relative clause

Preparation and characterisation of isotopically enriched[2] Ta_2O_5 targets for nuclear astrophysics studies

Abstract. The direct measurement of reaction cross-sections at astrophysical energies often requires the use of solid targets of known thickness, isotopic composition, and stoichiometry that are able to withstand high beam currents for extended periods of time. Here, we report on the production and characterisation of isotopically enriched[2] Ta2O5 targets for the study of proton-induced[2] reactions at the Laboratory for Underground Nuclear Astrophysics facility of the Laboratori Nazionali del Gran Sasso. The targets were prepared by anodisation of tantalum backings[3] in enriched[1] water (up to 66% in 17O and up to 96% in 18O). Special care was devoted to minimising[3] the presence of any contaminants that could induce unwanted[2] background reactions with the beam in the energy region of astrophysical interest. Results from target characterisation measurements are reported, and the conclusions for proton capture measurements with these targets are drawn.

1 Introduction

Stars spend most of their lives converting[3] hydrogen into helium through a sequence of reactions known[4] as the pp chain (for stars of mass $M \leq 1.5M\_$) or the CNO cycles (mainly for stars of mass $M > 1.5M\_$) [1,2]. An extensive experimental programme to study key reactions for hydrogen burning[5] in stars has been carried out over the last two decades at the Laboratory for Underground Nuclear Astrophysics (LUNA) facility in Italy [3,4]. Recently, the LUNA Collaboration has undertaken the study of proton-induced[2] reactions on 17O and 18O, important nucleosynthesis processes in several stellar sites, including[5] red giants, asymptotic giant branch (AGB) stars, massive stars, and classical novae [5,6]. In particular the ratio between the rates of the 17O(p, α)14N reaction (Q = 1191.8 keV) and the 17O(p, γ)18F reaction (Q = 5606.5 keV) affects the galactic abundance of 17O, the stellar production of the radioactive 18F nuclide, and the predicted[2] oxygen isotopic ratios in pre-solar grains [5]. For the study of such reactions, solid targets enriched[4] in 17O and 18O isotopes have been made.

[1] N.B. Be careful not to include words which are not active tenses! Remember many of the –ed and –ing forms are not active verb forms. The most common uses are: both –ed and –ing forms may be adjectives (2),–ing forms may be nouns (3), both –ed and –ing forms may introduce a reduced relative clause (4) and (5). Try to find examples of (4) and (5) in your texts.
[2] adjective
[3] noun form of the verb: when following a preposition, or when following a verb of pattern *verb + (do)ing*, or when it is the subject or complement of a verb. It may or may not be preceded by *a/an/the* etc. It may be plural (e.g. 'backings')
[4] reduced relative clause – past participle substitutes *which+ passive verb (any tense)*
[5] reduced relative clause – present participle substitutes *which+ active verb (any tense)*

Figure 2
Screen shot of a sample of digital tense analysis given to students as a model to follow.

3.3. Phase 3. Discovering how structural and rhetorical features of RA correlate with language choices

Classroom activities. The class was now ready to begin detailed discussion on the structure and discourse aspects of the RAs and to analyse the linguistic characteristics specific to each section. Using the same method as described above (alternating small group work and class feedback sessions), participants were guided through a series of concept questions to elicit a model for each section. They then compared the model with their individual corpora, noting structural and rhetorical aspects, such as evidence of the CARS introduction moves (Swales 1990) in the margins.

At least one two-hour session was dedicated to discussing each section and practising one or more related language points. By the end of this part of the course, working orally in small groups or pairs, the participants had explained the purpose of their work and a little of the theoretical background; they had described the method they had used in an experiment or a device they were involved in developing and they had described and interpreted a graph, or some other visual of their own research results. This was challenging work because although they all shared a common background knowledge of physics, they often knew little about each other's specialist fields.

4.Class corpus and related class activities

The 28 class corpus files were converted to .txt format by the teacher and loaded directly into AntConc (version 3.5.7; Anthony 2018), a freeware corpus analysis toolkit for concordancing and text analysis. The aim was to find examples illustrating difficult linguistic points chosen, as previously stated, on the basis of the author's professional experience. For the purposes of the course, time consuming "cleaning" of the files was deemed unnecessary. For this paper, two tricky words which often cause problems have been selected to illustrate the use that was made in class of the RA class corpus. The following paragraphs illustrate how activities were developed for illustrating and subsequently practicing an appropriate and correct use of "aim" and "allow".

4.1. Tricky word 1: "aim"

Clearly, as a word which expresses purpose, "aim" is most likely to be found in the introduction section of a RA, with possible use in the abstract and reiteration in the conclusion. These exercises and activities were therefore developed to use as part of the activities regarding the introduction section.

The difficulty many NNS find with this word is that they confuse the patterns which follow "aim" when it is used in the passive voice (meaning to

be directed towards something), as an active verb (meaning to plan/aspire/intend to do something) and as a noun (meaning target, intended outcome). Halliday (2004) would probably recognize metaphor processes at work in most phrases using this word, the original meaning presumably being that of pointing a weapon.

A concordance run for "aim*" (Figure 3), executed on the 28 files of the class corpus found 27 hits, three of which were ignored because AIM was used as a proper noun (the title of a computer programme), and was therefore judged to be misleading for the purposes of the class exercise. The author then imported the hits into a Word file for ease of display, and analysed the examples, classifying them according to the various correct or incorrect patterns (the latter were all produced by authors who are presumably NNS), and reducing the number of characters before and after the target. This produced phrases which were both sufficiently detailed and would allow students to focus their attention on the target phrase (Table 1, Annex).

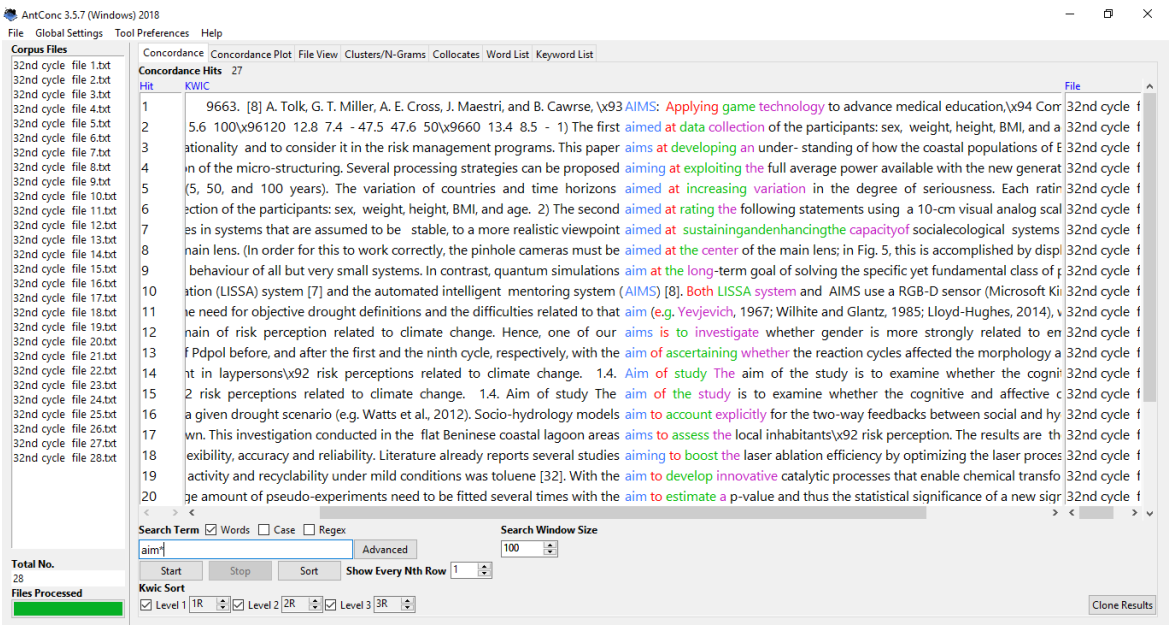


Figure 3
Screen shot of AntConc concordance results for "aim*"in the "dirty" 32nd cycle chosen RA class corpus. Hits 1, 10 and 27 were ignored because the word AIM was a proper noun. 5 examples with an incorrect pattern were found.

First the students were presented with (hopefully memorable) examples of the most common patterns. They were asked to imagine when or where they might hear/read the following and to paraphrase them:

- We aim to please, you aim too, please!
- Aim for the stars and you'll reach the sky!
- The aim of the game is to cite your own name!

- Why are so many viruses aimed at Windows? It crashes just fine on its own, thank you!

Next, the class was given two or three correct examples taken from each group in the table, emphasizing the pattern. A written exercise followed, in which students had first to identify the 6 out of 12 further examples from the corpus with an incorrect pattern and then try to find a correct way of expressing the idea. Finally, in pairs they asked and answered the following questions:

- What is the aim of your PhD project?
- What are you currently doing towards your PhD? What is this aimed at?
- Are you aiming to do an International English certificate?

4.2. Tricky word 2: "allow"

In the experience of the author, the word "allow" is presented repeatedly to students in the passive expression "(not) [be] allowed to" in non-specialist EFL course books, but rarely in the active form. As online English exercises appear to reflect the same emphasis, this statement can be quickly tested by a rough and ready, although hardly scientific, method. If we insert "allow, English exercises" into an internet search engine, the results will almost certainly produce a strong preponderance of "allowed to" exercises, despite our not having specified this in the search. This is fine, it is correct English and presumably the idea of permission/authorization is the most common meaning of "allow" in general English corpora, since it is reiterated so frequently in EFL text books. However, science writers are more likely to need to explain mechanisms where one thing allows another to happen, expressing relationships between objects and processes.

Hoey's claim that "grammar is the product of the accumulation of all the lexical primings of an individual's lifetime" (Hoey 2005, p. 159) could explain why, in the author's experience, both students and professional NNS science writers alike have trouble with "allow". Somehow, the juxtaposition of "allowed" and "to", which students have used and heard so frequently and which sounds so similar to "allow to", may be priming them to extend the same juxtaposition to the active pattern, thus producing phrases with the active verb "allow" directly followed by an infinitive instead of the required noun/pronoun complement. Alternatively, it could simply be a cross-language priming from their native language. It would be an interesting point to follow up, but is beyond the scope of this paper.

The concordance run for "allow*" (Figure 4) produced 75 hits, which were examined and organized by the same method as the "aim" hits. Examples, of the 3 main correct patterns, and ones which were incorrect and the awkward solution with "one" as the complement, were chosen from among the hits and

presented as “the Good, the Bad and the Ugly” (Figure 5). The students again worked on an exercise (Figure 6) in which some examples taken from the class corpus were correct and others required correction. In some cases, three alternatives were possible. It was necessary to add a few examples from the previous year because there were insufficient examples of the “allow + noun + to be/to be done” pattern in the 32nd cycle class corpus.

This exercise was followed by group work in which students challenged each other to describe different parts of a series of fantastical contraptions in Heath Robinson cartoons, such as The Deckeyrie For Unbalconied Flats[3] (Robinson, Browne 2015), answering the question: “what does this allow you to do?”.

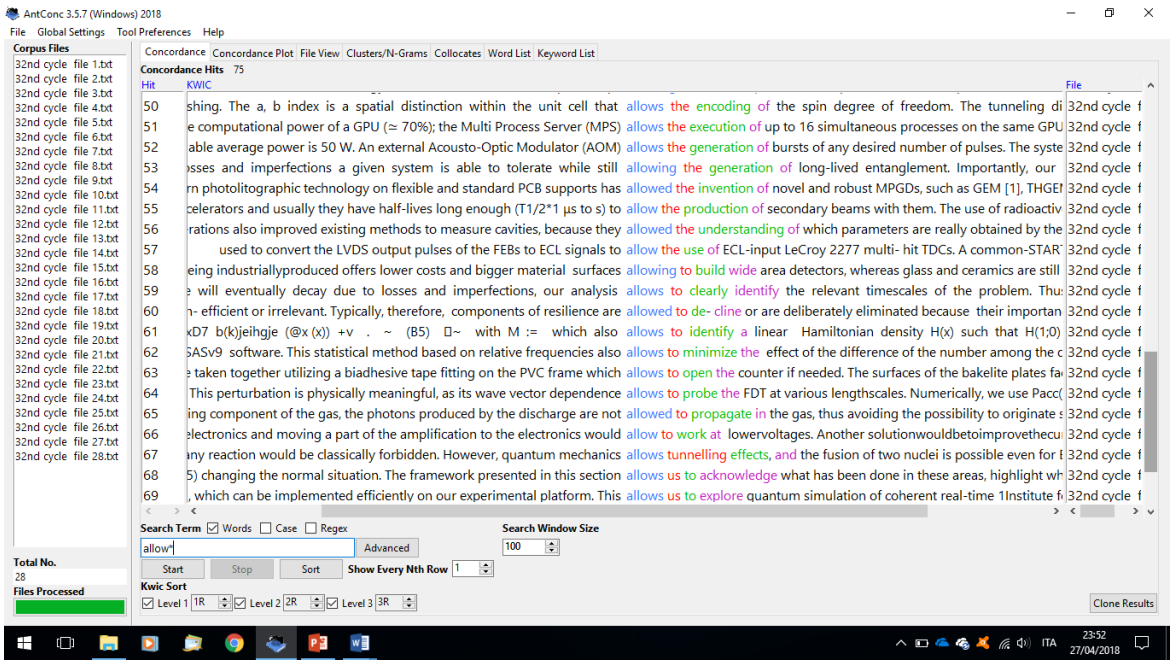


Figure 4
Screen shot of AntConc concordance results for “allow*” in the 32nd cycle chosen RA class corpus. Note hits number 60 and 65, the only two out of 75 hits to use the “(not) [be] allowed to do” pattern, and one used a similar pattern but in a reduced relative phrase. In the case of 65, “prevent” would probably have been preferable.

[3] Viewable at the following URL: https://www.spectator.co.uk/2017/05/there-was-method-to-the-madness-of-heath-robinsons-extraordinary-illustrations/. Accessed 27-04-2018.

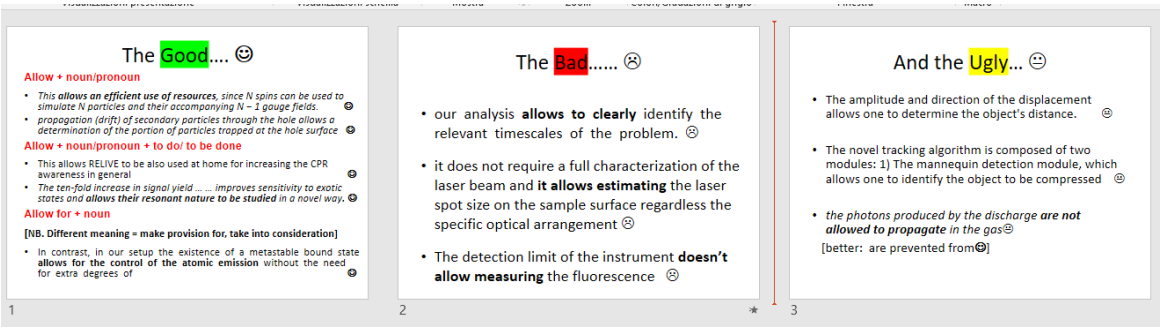


Figure 5
Screen shot of "allow" model and examples in ppt. slides shown to class during presentation phase of the exercise.

Exercise. Read the following extracts, most of which are from the class' chosen articles. In 5 of the examples the authors have used the verb correctly, in three it is correct but ugly, while in the remaining 5 they have got it wrong. Can you identify the 5 correct phrases? Now correct the others.

1. The hand detection module which allows one to allows one to identify and track the upper part of the rescuer\x92s overlapped...
2. it allows estimating the laser spot size on the sample surface regardless of the optical arrangement
3. Our platform allows direct measurements of the vacuum persistence amplitude and of the generated entanglement.
4. Brillouin microscopy can potentially allow more accurate determination of IOP
5. simple counting argument allows estimating the primordial 4He abundance
6. This perturbation is physically meaningful as its wave vector dependence allows to probe the FDT at various length scales.
7. the product of the two factors mentioned above allows the statistical uncertainty to be reduced.
8. III is devoted to the study of poles in the complex-energy plane, which allows us to extract crucial information relevant to the entanglement- by-relaxation protocol.
9. Furthermore, it introduces an important computational advantage, since it effectively allows one to split long simulations into a number of independent tranches, providing a natural and very efficient parallelization of the problem
10. the decrease of[sic] the time interval between consecutive pulses does not allow an efficient heat dissipation into the bulk material thus causing a temperature rise
11. This allows to reduce in a relevant way the amount of required memory in systems with multicomponent order parameters or in simulations of three-dimensional systems.
12. The panels are taken together utilizing a biadhesive tape fitting on the PVC frame which allows to open the counter if needed.
13. The amplitude and direction of the displacement allows one to determine the object's distance.

Figure 6
Screen shot of exercise discriminating between correct and incorrect phrases with "allow". The examples were taken from the class corpus, with the exception of a few examples from the previous year.

5. Conclusions

With this teaching project, the author set out to find ways of taking advantage of the benefits offered by individual specialist corpora in an EST context without asking the course participants to learn to use a concordancing tool. Acquiring greater awareness of, and ability to create, scientific discourse is crucial for young scientists in the first year of their PhD, which is an important transition stage between student and professional status. They have experience reading RAs but have not generally had the opportunity yet to consider in detail how a professional constructs a paper. The activities described above were integrated into the short course for first year physics PhD students with a view to offering students authentic, highly relevant materials as a resource for language learning in their specialist field. Student response to these activities was positive, the creation and use of personalized EST corpora helped to strengthen their motivation to dedicate time to language study because the language, rhetorical organization and content were perceived as highly relevant to their studies.

Some authors propose that students should learn to use a concordancing tool to become autonomous and responsible for their own continued progress (Lee, Swales 2006, p. 72). In the project presented here, instead, a choice was made to leave work with the concordancing tool kit exclusively in the hands of the teacher, given the tepid response of previous cohorts to the proposal of actually learning to use it themselves. This project has shown that it is feasible to use features of pdf readers in self-study activities to carry out rudimentary word searches in a limited specialist corpus relatively quickly, while simple manual highlighting on paper allow students to analyse linguistic aspects such as tense, modality, etc. Analysing overall RA structure and rhetorical functions lends itself better to work on paper, in the author's opinion, but could also be performed with margin comments in a pdf file. A pedagogical drawback remains that the implicit lack of commitment could mean that few will make the time to continue expanding their personal corpus after the course has finished. Clearly also, as some of the activities and classwork materials must be changed each year, it is a relatively time-consuming process for the teacher, as Charles (2012) also notes. Each individual teacher must decide if it is worth the effort.

A technical drawback of using a personalized corpus is that the amount of text is necessarily very limited, compared with large corpora. In fact, even using the class corpus, insufficient examples of some patterns were found. Nonetheless, the author found it was possible to integrate them with examples from previous years.

In planning a classroom project integrating work on corpora, EST teachers also need to take into consideration the impact of financial, bureaucratic and logistic aspects. Milizia (2013) describes giving students access to a large political corpus using WordSmith, although she does not mention if this is the

freeware version, while Adams (2006) and Lee and Swales (2006) use both large and personalized corpora. Projects which involve access to precompiled corpora and/or paid software may have heavy costs (e.g. licenses to be made available to students and teachers, for use on university computers and/or at home) and involve bureaucratic and logistic challenges (e.g. time spent liaising with administrators and laboratory availability). Instead, the use of simple self-compiled corpora and freeware such as AntConc, as proposed here and in Charles' (2012) project, confer the advantage of ease of access both to the corpora and to the concordancing tools, thus reducing costs and simplifying bureaucratic/logistic issues.

Another unresolved issue in common with projects based on authentic discourse, raised here by the results of the concordancing of "aim" and "allow", is that of which linguistic model to choose to present to students. As can be seen from the percentage of errors found in this small class corpus consisting of highly specialized RAs, poorly written English sentences regularly slip through publishing vetting processes. The aim of this paper is not to address in depth the interesting question of whether NS models must always be preferred, or whether, considering that these students are using English as a Lingua Franca, we should accept any forms commonly used by NNS. However, we do need to point out to our students that at least three major aspects of their work will be evaluated when they submit an abstract or an article to a journal, or make a presentation: first and foremost its scientific content, but also their ability to express themselves in correct grammatical English and their knowledge of the unwritten rules of scientific discourse, a good command of which distinguishes an expert from a novice and will play an important role in helping them to pass the gatekeepers of their discipline.

In conclusion, as the author hopes to have shown, these three aspects are intimately interwoven and therefore need to be analysed and practised together, at specialist level. Much can be achieved using individually chosen specialist corpora in an EST class by a creative teacher, without asking students to learn to use concordancing software, thus providing aspiring scientists with exactly the ZPD material they need.

Bionote: Carmela M. White, 'CEL' and contract professor at Bari University, has worked in EST since 1985 and is currently attached to the Physics Department, where she is responsible for physics and materials science undergraduate and mathematics postgraduate courses, working closely with departmental staff to integrate language courses with the students' mainstream studies. She has presented several papers at the TESOL Italy annual conference on topics ranging from EST course and material development to motivation studies and tackling specific language difficulties faced by science students. The ideas presented here have been developed whilst teaching short courses for science PhD schools.

Author's address: carmelamary.white@uniba.it

References

Adams R. 2006, *Developing professional phraseology: a corpus linguistics approach*, in Mickan P., Petrescu I. and Timoney J. (eds.), *Social practices, pedagogy and language use: studies in socialisation*, Lythrum Press, Adelaide, pp 72-82.

Anthony L. 2018, AntConc (Version 3.5.7) [Computer Software], Waseda University, Tokyo, Japan. http://www.laurenceanthony.net/software

Charles M. 2012, *'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building*, in "English for Specific Purposes" 31 [2], pp. 93-102.

Dressen-Hammouda D. 2008, *From novice to disciplinary expert: Disciplinary identity and genre mastery*, in "English for Specific Purposes" 27 [2], pp. 233-252.

Doumont J. 2010 (ed.), *English Communication for Scientists*, NPG Education, Cambridge, MA. https://www.nature.com/scitable/ebooks/english-communication-for-scientists-14053993 (30.04.2018).

Gross A.G., Harmon J.E., Reidy M.S. 2002 *Communicating Science: The Scientific Article from the 17th Century to the Present*, Parlor Press, Anderson, SC.

Halliday M.A.K. 2004, *The Language of Science: Volume 5*, in Webster J.J. (ed.), *The collected works of M.A.K. Halliday*, Continuum, London.

Hoey M. 2005, *Lexical Priming: A new theory of words and language*, Routledge, London/New York.

Hyland K. 2008, *As can be seen: Lexical bundles and disciplinary variation*, in "English for Specific Purposes" 27, pp. 4-21.

Hyland K. 2009, *Academic Discourse*, Continuum, London.

Hyland K. and Tse P. 2004, *Metadiscourse in academic writing a reappraisal*, in "Applied Linguistics" 25 [2], pp. 156-177.

Krashen S.D. 1987, *Principles and practice in Second language Acquisition*, Pergamon Press Inc., New York.

Kress G. and van Leewen T. 1996, *Reading Images. The Grammar of Visual Design*, Routledge, London/New York.

Lantolf J.P. and Poehner M.E. 2014, *Sociocultural Theory and the Pedagogical Imperative in L2 Education: Vygotskian Praxis and the Research/Practice Divide*, Routledge, London/New York.

Lee D. and Swales J. 2006, *A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora*, in "English for Specific Purposes" 25 [1], pp. 56-75.

Mariotti C. 2012, *Genre Variation in Academic Spoken English: the Case of Lectures and Research Conference Presentations*, in Maci S.M. and Sala M. (eds), *Genre Variation in Academic Communication: CERLIS Series Volume 1*, CELSB Libreria Universitaria, Bergamo, pp. 63-84.

Milizia D. 2013, *Phrasal Verbs and Phrasal Units: Political Corpora within the Walls of the Classroom*, in Desoutter C., Heller D. and Sala M. (eds.), *Corpora in specialized communication: CERLIS Series Volume 4*, CELSB Libreria Universitaria, Bergamo, pp. 135-164.

McEnery T. and Xiao R. 2010, *What corpora can offer in language teaching and learning*, in Hinkel E. (ed.), *Handbook of Research in Second Language Teaching and Learning* Vol. 2, Routledge London/New York, pp. 364-380.

Robinson W.H. and Browne K.R.G. 2015, *How to Live in a Flat*, Bodleian Library, Oxford (first published 1936).

Spagnolo V. et al. 2013, *THz quantum cascade laser-based quartz enhanced photo-acoustic sensor*, in "38th International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz)", pp. 1-3.

Swales J. 1990, *Genre Analysis*, Cambridge University Press, Cambridge.

Tarantino M. 2004, *Epistemic and dialectic pathway to knowledge, meaning and language Advancement*, in "LSP & Professional Communication" 4 [1], pp. 69-88.

Tarantino M. 2011, *Modality categories in multimedia genres*, in "Rassegna Italiana di Linguistica Applicata" 3, pp. 157-183.

Wenger E., McDermott R. and Snyder W.M. 2002, *Cultivating Communities of Practice: A Guide to Managing Knowledge*, Harvard Business School Press, Boston, MA.

White C.M., forthcoming, *EST or CLIL? Integrating language courses within science degree programmes,* accepted for publication in "PERSPECTIVES A Journal of TESOL Italy".

Widdowson H.G. 1979, *Explorations in Applied Linguistics*, Oxford University Press, Oxford.

Annexes

Hit n°	Beginning of phrase	Target and completion of phrase	Text N°
	a. "Aim" used correctly as noun (+ of), or as noun [be] + infinitive		
21	the system was adopted for two different	aims: 1) to estimate the accuracy of the hand tracking	24
11	the difficulties related to that	aim, which we will not repeat here	18
14	1.4.	Aim of study.	8
15	The	aim of the study is to examine whether the cognitive and affective components of risk	28
13	with the	aim of ascertaining whether the reaction cycles affected the morphology and the dispersion	3
12	Hence, one of our	aims is to investigate whether gender is more strongly related to emotional reactions	28
	b. Verb "aim" use correctly with at + noun (hit 9 only), or passive verb [be] aimed at +-ing		
9	In contrast, quantum simulations	aim at the long-term goal of solving the specific yet fundamental class of problems	19
8	the pinhole cameras must be	aimed at the center of the main lens	4
2	1) The first	aimed at data collection of the participants	24
6	2) The second	aimed at rating the following statements using a 10-cm visual analog scale	24
7	a more realistic viewpoint	aimed at sustaining and enhancing the capacity of social ecological systems to adapt to uncertainty	26
5	The variation of countries and time horizons	aimed at increasing variation in the degree of seriousness.	28
c.	Verb "aim" used correctly active + infinitive		
18	Literature already reports several studies	aiming to boost the laser ablation efficiency by optimizing the laser process parameters	8
25	resulting in long-term adaptations	aiming to reduce impacts of drought in the future	18
16	Socio-hydrology models	aim to account explicitly for the two-way feedbacks between social and hydrological processes	18
17	This investigation conducted in the flat Beninese coastal lagoon areas	aims to assess the local inhabitants risk perception.	27
22	great potential for use in drought research is comparative analysis, which	aims to find patterns by analysing a large set of catchments with a wide range	18
23	RELIVE	aims to go beyond the state of the art accurately estimating the CC depth	24
d.	"aim" used incorrectly		
19	With the	aim to develop innovative catalytic processes that enable chemical transformations to be performed	3
20	pseudo-experiments need to be fitted several times with the	aim to estimate a p-value and thus the statistical significance of a new signal	11
24	to study the entire interrelated system with the	aim to put the pieces of the puzzle together.	18

26	focus on a specific point or research question with the	aim to solve part of the puzzle, or to study the entire interrelated system	18
4	Several processing strategies can be proposed	aiming at exploiting the full average power available with the new generation of ultrafast laser	8
3	This paper	aims at developing an understanding of how the coastal populations of Benin perceive natural	27

Table 1
24 hits of “aim” organised according to pattern, including 6 incorrect examples.