Encuentro
Revista del Departamento de Filología Moderna

GUIDELINES FOR THE DESIGN AND GENERATION OF CORPUS-BASED TERMINOLOGY-ORIENTED INSTRUCTIONAL MATERIAL

DIRECTRICES PARA EL DISEÑO Y LA GENERACIÓN DE MATERIAL DIDÁCTICO BASADO EN CORPUS PARA EL APRENDIZAJE DE TERMINOLOGÍA

Araceli Losey León
Universidad de Cádiz

Abstract

A great deal of research has been conducted to exploit linguistic corpus data for language learning and teaching use. In the integrative corpora and language pedagogy stream of research diverse practical purposes coexist at the front and have presently become consolidated practice -learner corpora, ELT dictionary-making corpora, learner dictionary corpora, EAP corpora, specialised domain corpora, English-taught programmes corpora such as Content Language Integrated Learning (CLIL) or English-medium Instruction (EMI). This paper outlines a framework for corpus exploitation process towards the design of instructional material for the learning of terminology of a specialised domain which is applicable to ESP courses. Our proposal advocates for a protocolised use of written corpora to devise a resource for teaching terminology and, accordingly, suggests guidelines to ESP teachers/ practitioners who undertake the tasks of extracting corpus data and adopt the role of material designers and, occasionally, of computer programme developers. To accomplish this, in the first place, specialised corpus compilation and design criteria have been customised to guarantee consistency across the pedagogic ends, the functional language needs of the ESP user and the specific learning objects within the DDL approach framework. In the second place, search strings used for semi-automatic text mining actions are illustrated as well as samples of terminology-oriented activities that can be automatically generated using concordancers and annotating computational tools. Future research might apply our proposal to further language contexts.

Key Words: Didactic exploitation of specialised corpora, ESP innovative educational instruction, Natural Language Processing, automatised generation of terminology-oriented learning material, Language Learning and Technology

Resumen

Numerosas investigaciones se han llevado a cabo para explotar los datos de los corpus lingüísticos con fines de enseñanza y aprendizaje. En las líneas de investigación sobre la integración del corpus en la didáctica de la lengua inglesa, son varias las parcelas que coexisten en posiciones destacadas y que, con la práctica, han terminado consolidándose –corpus lingüísticos de estudiantes, corpus de enseñanza de lenguas para la elaboración de diccionarios, corpus para diccionarios de estudiantes, corpus de inglés para fines académicos, corpus de inglés para fines específicos, corpus para aprendizaje de contenidos a través del inglés como el Aprendizaje Integrado

de Contenidos y Lenguas Extranjeras (AICLE) o la Enseñanza por medio del Inglés (EMI)-. En este artículo se delinea un marco desde el que desarrollar el proceso de explotación de corpus para diseñar material didáctico orientado al aprendizaje de la terminología de un dominio especializado y aplicable a cursos de IFE. La presente propuesta aboga por un uso protocolizado de los corpus escritos para desarrollar un recurso de enseñanza de la terminología y, por consiguiente, sugiere un conjunto de directrices a los profesores de IFE que acometen las tareas de extracción de datos del corpus y adoptan el rol de diseñadores e, incluso, de desarrolladores de programas informáticos. Para ello, en primer lugar, este estudio plantea un ajuste de los criterios básicos de compilación y diseño de corpus especializados a los fines pedagógicos, a las necesidades funcionales y lingüísticas del usuario de IFE y a los objetivos específicos de aprendizaje, entre otros, dentro del marco del enfoque de aprendizaje basado en datos. En segundo lugar, se ilustran las opciones de búsqueda empleadas en las acciones de minería de textos así como muestras de actividades orientadas a la terminología que pueden ser generadas automáticamente a partir de las herramientas de concordancia y los anotadores. Futuras investigaciones podrían extender esta propuesta a otros contextos del lenguaje.
Palabras clave: Explotación didáctica de corpus especializados, innovación educativa en IFE, Procesamiento de Lenguaje Natural, generación automatizada de material didáctico para el aprendizaje de la terminología, Tecnología para el Aprendizaje de Idiomas

1. INTRODUCTION

Direct and indirect uses of specialised language corpora for language/content instruction have opened up new strands of linguistic inquiry and have boosted empirical investigations in a tradition which is now firmly established in the LSP context. Specialised language corpora have been dramatically exploited in the last 30 years for an extensive variety of descriptive and applied purposes and across theoretical areas of all kinds. Owing to the corpus interplay, it is usual to find hybrid approaches to Terminology, Translation, ELT, Lexicography and ESP studies, to name but a few. Another proof if the corpus evolution is its omnipresence. Corpus use is no longer confined to the fields of Language and Humanities but it is extensive to non-linguistic disciplines (Pharmacology, Law, Engineering, Maritime Navigation…).

Looking back, one of the earliest corpus-based English language instructional materials was Willis & Willis's Collins COBUILD English Course (1988), which was based on the same corpus research that produced the Sinclair's pioneering *Collins Cobuild English Language Dictionary* in 1987. While educational technology continued its progress, corpus management tools kept at pace and gradually entered into further fields, finally widening its scope to ESP. The corpus paradigm, as Laviosa (1988) referred to, is embedded in educational technology and Terminology is one of the disciplines which have benefited the most from its potential in the research and teaching duality. On the one hand, corpus prevails as a methodological tool to empirically uncover the term's communicative, conceptual and linguistic dimensions. On the other hand, corpus invigorated vocabulary teaching and hence its pedagogical value. As Boulton & Pérez-Paredes (2014) put forward the interest relies now on how to integrate corpus linguistics tools and techniques in language pedagogy.

In this regard, we can cite preceding scholarly work addressing this issue. Among them, we can mention Luzón-Marco (2000), in the Medical English field, Curado-Fuentes (2002), in the Business

English field, Marinov (2013), in an undergraduate course of English for Tourism, Wu (2014), who proposed direct use of corpus to ESP students of Technical College in Taiwan, and Marzá (2014) who developed corpus-based DDL research in English for Tourism field. Further insights transferred the corpus use to genre-based writing pedagogy (Mizumoto, Hamatani & Imao 2017; Chang 2014).

Even though advantages of applying corpora to ESP pedagogy have been heavily discussed some issues remain to be addressed. In this vein, Chirobocea (2017, 364) explains that devising corpus-based ESP instructional material method "is still largely either misapplied or misunderstood by teachers and, apart from its obvious advantages, it has some important disadvantages which make it harder to use". In our view, the development of corpora for materials of an ESP programme ought to be regulated to ensure better learning of the targeted specific vocabulary.

The present work intends to contribute to ease the corpus applicability to ESP by following a protocolised use of corpus and tailored guidelines for compilation criteria and design adapted to ESP aims and user ends. To this end, the study is structured as follows. Firstly, it shall be provided a full account of corpus design criteria for specialised domain and a compilation methodology aimed to serve as a guide for ESP teachers. Secondly, we will move on to illustrate how to exploit the corpus bearing in mind the terminology instructional goal. Thirdly, a workflow of actions for the corpus processing by basic computational tools is proposed. Fourthly, samples of automatic generation of terminology oriented activities shall be presented. To finish with, we shall present the advantages and drawbacks that have been encountered along the research.

2. THEORETICAL FRAMEWORK

The advent of corpora and educational technology has paved the way to unfold, to discover, to describe, to verify and to evaluate language uses in particular domains, introducing new empirical ways to the qualitative and quantitative studies of language, its structure and patterns. Basically, our study is primarily confined to extract the corpus data containing the terms that, included in the content of a teaching lesson, the learner is to practise. Certain pedagogical approaches have found their way into the corpus-based studies. One of them is the implementation of the lexical approach (Lewis 1997) where the instruction content is designed around frequent vocabulary and uses the findings in a corpus. Data Driven Learning or DDL (Johns 1991) is another corpus-based approach characterised by conferring the corpus direct contact to the learner either as a user (receptive role) or as a compiler and analyst (producer role). Our pedagogical assumptions rely on the potentialities of learning through a term's contextual environment, on the DDL approach as facilitator and on Corpus Linguistics as the purposeful textual inputs to study language using computer processing tools and a set of quantitative and qualitative procedural means and methods, to put it simply. We can also say that a context-based learning approach (CBL) is at the core of our proposal. The sound role of contextual information that a corpus provides is essential to grasp the conceptual and linguistic information of a terminological unit. A term environment can help to identify its meaning when seen in relation to others. In addition, ESP learners can access and learn through samples and models of real language within an authentic specialised communicative setting. The typical learning tasks and activities associated to CBL are scanning, exploring, selecting,

assembling, constructing, mapping, and sequencing. On the other hand, CBL foremost intends to generate positive effect on attitudinal and cognitive aspects of learning. The former one is achieved in terms of learner's sustained motivation, autonomy and engagement with authentic texts in an educational and technological setting and the latter one enables the upgrade of the learner's term conceptualization as well as memory enhancement.

Furthermore, all these views are in balance with the linguistic approach to terminology represented by the Communicative Theory of Terminology (Cabré 1999), probably the most currently backed up approach and adopted in this study. It puts emphasis on the communicative aspects of the use of the terms, takes cognitive and discursive aspects into account and the object of study are terms conceived in the context of specialised discourse that are actually used in LSP corpora. Thus, this theoretical framework goes along with the learning object of the corpus-based instructional material that we seek to develop.

We present one way of exploring corpus possibilities for teaching but there are others. Our proposal encompasses mixed functional and content-based syllabi where corpus-based terminology instructional material is to be used as sequential part of a didactic unit or as the source for addressing further (extra-) language and discursive features

3. DATA DRIVEN LEARNING AND ESP CORPUS-INFORMED INSTRUCTIONAL MATERIAL

The data-driven learning is an inductive learning that helps the learners to uncover the language behaviour on their own. Previous studies in the language teaching tradition have addressed this issue and have illustrated its applicability (Leech 1997, Gabrielatos 2005, Scott & Tribble 2006, Campoy, Gea-valor & Belles-Fortuno 2010, Flowerdew 2012, to name but a few). In this section, the material developmental stages are described together with the DDL facets which were applied. The ESP teacher/practitioner prevailing requirements for monitoring all the process are outlined as well.

3.1. DDL modes

A bottom up approach shall be conducted to exploit the corpus because corpus data are our starting point to build up learning material. Terms will not be extracted according to prior inductive proposal but the learners undertake (semi-) guided (un-) controlled discovery tasks to get to the term keyness. According to Leech (1997) there are different modes of approaching Data Driven Learning. Two of the most typically used in tertiary education are:

-A 'teaching to exploit' DDL approach: In this mode, the ESP learners compile the corpus and work 'hands-on' being semi-guided by the teacher. We refer to this mode as 'open DDL'.

-A 'exploiting to teach' DDL approach: the ESP learners handle the corpus under the ESP teacher controlled actions, that is, the ESP teacher is previously trained in corpus compilation and processing, devises corpus-based tasks and presents them to the learner in a printed format or embedded in the corpus text. We refer to this mode as "controlled DDL".

What we adopt is a data-driven approach where both of them, the open and the controlled modes, can be applicable though dependant on the learners' age, level and the time constraints. We set out to develop instructional material driven by data obtained from a specialised corpus which has been particularly compiled to this end by the teacher who presents the material to the learner.

The learning activity proposed consists of two parts that shall be sampled in section 4.3, a data-driven learning action (pre-warm corpus-based activity oriented to the learner's observation and awareness of the term behaviour) and a data-driven activity (a specific corpus-based task on terminology learning aimed at term-identifying and usage tasks).

Nevertheless, we are aware of the fact that the controlled DDL approach is easier to be organised and performed during ESP lessons. The learners would then manage a corpus that has been previously compiled and designed by the teacher and search the query on his/her own following the ESP teacher's instructions. They work either on a corpus building/ processing tool. The resulting activity is learner-centered as well and, according to Bernardini (2000), it can enhance memory retention and recall because learning through data involves a high degree of task involvement.

3.2. The ESP Teacher engaged in corpus-informed DDL

Even though there are corpora available on the web, we propose that ESP teachers build their own corpus to ensure that the corpus-based learning activities meet the pedagogical needs and are suitable for the final aims and learning objectives. Some preliminary actions are expected to be carried out by the ESP teacher involved in the design and development of a corpus-informed DDL instructional material, namely:

(1) The ESP teacher should, alone or in collaboration with field's experts, make some decisions on the corpus domain under study. Its selection has to be relevant to the specialised language needs of the ESP learner. It entails a thorough revision of the topic contents of non-linguistic courses and consultation with experts on the conceptual fields and the terminology typically associated before determining which terms to be learnt.

(2) ESP teacher's role as a corpus-based learning material designer requires the acquisition of corpus computer management skills.

(3) ESP teacher's role as researcher to distil the key terms (monolexematic and polylexematic units) in context.

(4) ESP teacher should be advised by the field's experts on the document text types which are prototypical in the domain and useful for specialised communication as well as on the validation of terms' process.

(5) The ESP teacher should also make decisions on the scope and range of the corpus size.

(6) The ESP teacher also needs necessary skills to interpret quantitative data, especially evidence obtained from computer tools such as Wordsmiths tools or AntConc.

(7) The ESP teacher has to check access to technical resources (corpus processing tools).

(8) The ESP teacher may provide the learners with further samples of lexicogrammar patterns, despite the focus is on terminology, so that they can perceive that the terms are not seen in isolation but in context.

(9) The ESP teacher is recommended not to strive for overly ambitious targets but for a balanced and rational scope in the learning object.

Despite the fact that most teachers seem to be resistant to integrating corpus in their syllabi, the ESP teacher tends to be particularly engaged in creating material, especially in those fields with a scarcity of pedagogically oriented materials.

3.3. Instructional material devise: stages and follow-up

The first stage in the material design consists of the documentary selection planning before corpus compilation takes place in a second stage. The corpus processing operations are performed at a third stage. Next, once the term candidates' extraction and the term validation have been executed, the learning activity is generated and fulfilled by the learner at a fourth stage. The full process ends up with the fifth stage focused on assessing the learner's performance, reviewing, filtering and, if necessary, refining the learning activity. Table 1 below shows the working stages and the follow-up.

Stages	Actions proposed to be developed in the open DDL approach	Actions proposed to be developed in the controlled DDL approach
FIRS STAGE: PLANNING AND TIMING	-Setting the learning objective and assessment scale. -Topic selection. -Timing of sequences of actions. -Verifying availability of computer tools. -Creation of a term validation template. -Development of instructions for learners on corpus compilation and on corpus processing. -Guiding the learners to the terms in focus. -Development of a post-activity survey.	-Setting the learning objective and assessment scale. -Topic selection. -Timing of sequences of actions. -Verifying availability of computer tools. -Creation of a term validation template. -Development of a post-activity survey.
SECOND STAGE: CORPUS BUILDING (Using dedicated guidelines)	- The ESP teacher and the ESP learner select electronic documentary sources in English. -Compilation of an ad hoc specialised English corpus by the learners.	-The ESP teacher selects electronic documentary sources in English (monographs, textbooks, research articles, handbooks, manuals, press articles, regulations on the field …). -Compilation of an ad hoc specialised English corpus by the ESP teacher.

THIRD STAGE: PROCESSING (Using dedicated guidelines)	-Using computational tools to analyse texts. -The learner uses the instructions for corpus processing which have been developed in the first stage. -The learners follow the instructions on the terms to be extracted. -Term candidates' extraction is made by the learner. -Term validation by the learners (field's semi-experts) under the guidance of the ESP teacher using the validation table designed in the first stage.	-The ESP teacher is trained on managing software tools. -The ESP teacher designs the corpus. -The ESP teacher compiles the corpus. -The ESP teacher processes the corpus. -The ESP teacher extracts the candidate terms -The ESP teacher validates the terms in cooperation with the subject-matter expert. -The ESP teacher extracts the data and transforms them into information that will be shaped as instructional material oriented to terminology learning.
FOURTH STAGE: Performance	-The learner stores the data for glossary development. -The learner fulfils the tasks.	-The ESP teacher presents the corpus preview to the learners. -The ESP teacher presents the activities which have been designed using the data extracted from the corpus. The learning activities/ tasks can be hosted in a programme (Microsoft Access…) or presented on a printed format. -The learner fulfils the tasks.
FIFTH STAGE: Analysis, evaluation and refinement	-The learner's tasks are evaluated. -The learner completes the post-activity survey. -The ESP teacher analyses the results obtained.	-The learner's tasks are evaluated. -The learner completes the post-activity survey. -The ESP teacher analyses the results obtained.

Table 1. Stages in the production and follow-up of the corpus-based ESP instructional material

It is worth mentioning that the key terms are based on the criteria of frequency measurement so that we start from the assumption that term frequency is an indicator of its relevance and outstanding role in a field of knowledge. It is usual to resort to frequency for justifying the introduction of these key terms also known as KWIC (key word in context) in the teaching and learning lesson plan.

4. GUIDELINES FOR THE DEVELOPMENT OF AN AD HOC WRITTEN SPECIALISED LANGUAGE CORPUS

Even though corpus building criteria are seemingly applicable to language teaching, we advocate for particular criteria for dealing with ESP teaching material. ESP requires the application of specialised language criteria for corpus compilation in consonance with the ESP learners' needs. The specialised teaching material should primarily feature ways to promote the learners' communicative competence and the ESP teacher should ascertain and control that the material is pedagogically relevant to fit the real specific terminological needs of the learner.

Building a written/oral corpus requires the use of guidelines since a corpus is not a simple gathering of texts. Selecting, discarding, filtering, structuring, classifying and categorizing are essential activities to prepare a corpus and recall its data at will before the processing comes into play. In this section we are concerned with providing simple guidelines to teachers who have not already attempted to manage corpus and wish to do so in order to transfer the data as pedagogical resources into their courses. Although our study corpus belongs to the domain of Science & Technology, the orientations are not confined to but extensive to any area in the Humanities and Social Sciences' disciplines, among others. The general schemata for corpus building are presented below. All of them are external criteria following those proposed by Atkins, Clear and Ostler (1992), Bowker and Pearson (2000), Corpas (2001), Losey-León (2015) and Seghiri (2017).

4.1. Design criteria

a. Specifying the aim: Clarifying the objective of the text compilation is at the core of corpus design. It can be oriented to a product, namely:

-terminographic product: glossary, thesaurus, (term-)ontology,…

-translation memory,

-knowledge data bases,

-language instruction material (for translation training, ESP teaching, CLIL/EMI teaching).

or centered on extracting data to analyse the common pattern and vocabulary and draw conclusions other than those derived from a terminographic product:

-Language analysis and description to gain knowledge about specialised language: typical activities are describing and comparing the language of the domains and subdomains, contrasting terminology and uses, discovering neologisms, analysing phraseology, extracting terminological collocations (Costa & Silva 2004), analysing pattern frequencies, identifying variation cases, (post-) editing and annotating,

-Error detection and analysis.

b. Clear target-user: It involves guessing about the tasks the user has to tackle successfully (ESP language learner, specialised translator, researcher,…)

c. Deciding on the time period coverage: the corpus can be synchronic (providing data from one point in time) or diachronic (dealing with the way language develops over time). So, the choice depends on whether the compiling user is interested in language across the years or at a definite time. Both of them are useful to study language variations and neology.

d. Selecting the language and the number of languages: corpus can include texts in one language (monolingual corpus), two languages (bilingual corpora), or more (multilingual corpora).

e. Degree of relationship between the languages: parallel corpora (segments of source texts that have been aligned with their equivalent segments in the target language) or comparable corpora (when the documents are compiled in two or more languages but there is no exact equivalence between them).

f. Organising the corpus samples into different conceptual fields. It is advisable to make further subdivision in genres.

g. Degree of specialisation of the documents contained in the corpus. A research article provides specialised information of a domain and is expected to be read by (semi-) experts in the field. For instance, a leaflet on instructions to evacuate a site is expected to be read by laymen, in broad terms. It is crucial to make decisions on this feature which is determined by the learners' mastery of English and the suitability of the text type. A corpus may contain a single text type or a combination of different text types. They should be stored in separate files as it enlarges the scope and range of future analysis. This feature has a direct influence on corpus balance issues.

h. Establishing modularity capacity (Losey-León 2015): It depends on whether your aim is to collect information on a broad domain or on partitions (subdomains, text types, …). It allows you to study the corpus as a whole or in sections being a fertile land for contrastive analysis.

i. Authenticity: This is a thorny issue and there are conflicting views. Some studies argue that authenticity in language pedagogy should be relatively aimed and texts should be adapted, if necessary. However, in our view, texts should reflect real unabridged language when dealing with specialised domains. In this line, Wu (2014, 123) pointed out that re-writing techniques such as paraphrasing might obscure and alter the integral essence of the text so he concluded that one solution to guide the readers through a text could be the use of various devices, without making drastic changes to the content.

j. Sample size: There is not an absolute criterion on corpus size but general language corpora are, by far, of larger size than specialised language corpora. On the other hand, as regards specialised corpora, vast amount of texts do not necessarily correspond to a substantial body of expertise (Losey-León 2015, 296).

k. Balance: It features the amount of information provided for corpora containing a wide heterogeneity of text types which poses the problem of achieving accurate terminological portray of the domain. If the dataset is biased, when submitted to quantitative and qualitative study, the results are not reliable.

l. Representativeness: As Biber (2003, 256) stated, the design of a representative corpus is not really finalised until the corpus is completed and the parameters of variation are obtained and analysed. As a matter of fact, a corpus can be deemed representative as far as it allows completeness according to the

learning objectives in an educational context. This issue has been dealt with extensively by Corpas and Seghiri (2007).

4.2. Workflow of corpus compilation, term extraction and processing

Once the corpus design has been completed, the empirical study begins. Figure 1 synthesizes all the central tasks ranging from data gathering to corpus processing and results.

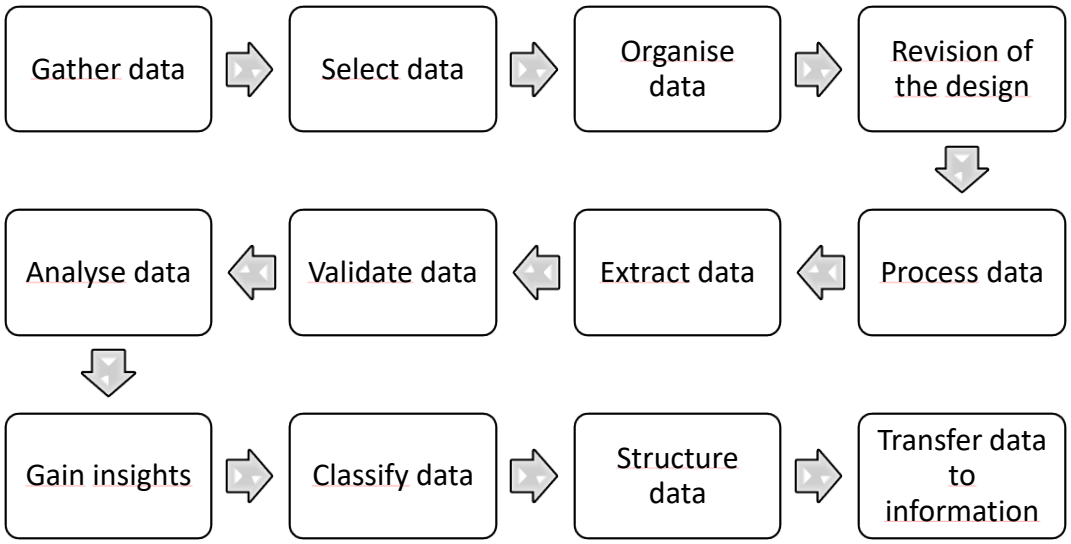

Figure 1. Workflow diagram of specialised corpus management tasks from data to information transfer

1) Data gathering. This is the compilation and documentary stage. It consists of the following set of actions:
    a. Consultation of specialised sources (electronic or printed format):
        i. Company organisations dealing with the (sub)domain.
        ii. Documentation from related institutions and organisations (standards, technical documentation, …)
        iii. Subject-matter textbooks
        iv. Subject-matter research articles, periodicals, reports…
    b. Consultation of terminographic sources
    c. Consultation of databases, termbases, …
2) Data selection. This is the set of documents that will be included in the corpus. Decisions on the text type, level of specialisation and time span have already been made.
3) Data organisation. Its classification arrangement is essential for the accessibility of the specific documents when queried. At this point, it may be necessary to check the files and convert them into txt format which is the most extensively required by computer processing tools. Apart from this, the files in the original format should be kept and stored. The files should be categorised

according to text type and then subcategorised according to the language, conceptual field, period or the other way around. Expert advice is recommended. Tags or codes accompanying the files should contain distinguishable features to ease the location. Further actions at this time are the storage of the corpus in a single file and tag it as reference corpus; then, select one of the folders as the study corpus. This action will allow the ESP teacher to compare and refine results if the aim is to get the finest specific terms of a (sub) domain.

4) Revision. This is the moment to check the adequacy of the corpus compiled, of its content, of its languages and of its size to our aims.
5) Process data. It requires a previous preparation of a stopword list in the language concerned to filter and discard unwanted signs, symbols and noise. This is stored in a separate file. It will be available for filtering the preliminary results. As regards the data processing, reliable software is available on the web and provided with tools (wordlist, concordance, keyword, clusters, N-grams…) that shall be used according to the ESP teacher inquiries. The processing method is frequency based. Some well-known programmes are:
    a. Wordsmith's tools 7.0 (Scott, 2019 - late version)
    b. AntConc (Anthony, 2017 - late version)
    c. Sketch Engine
6) Extract data. It is recommended to start using the wordlist tool. It will provide the teacher with the first quantitative data of the corpus such as nº of tokens, nº of types and the tokens/type ratio. This indicator gives the ESP teacher information about the lexical density of the corpus. Most tools allow uploading of a stopword file. This action is recommended to obtain finer results. Since the interest relies on terms, we can then apply the concordance tool so that we can obtain all the instances in which a word co-occurs with others. The results will yield collocational and lexicogrammar patterns. We can now apply the keyword list tool that will provide us with a term candidate list according to statistical criteria and parameters provided by the programme. It is essential to be aware of the fact that these lists contain candidates to terms because their validity has to be tested. The user can also ask the computer to search the corpus for strings of words. This technique is associated to Biber et al. (1998).
7) Validate data. At this stage the cooperation with the subject-matter expert's opinion is essential. We can have a template ready to facilitate them the task. The final list will contain the terms. The qualitative results obtained complement the quantitative results yielded in previous stages.
8) Analyse data. The terms can be now displayed in context and analyse in terms of relevance.
9) Gain insights. The ESP teacher can evaluate which terms in context can be practised from the corpus preview.
10) Classify data. From the data obtained it is possible to categorise, make generalisations, and group data. At this point, it is usual to distinguish special features and envisage the design of an activity.
11) Structure data. The data obtained should be structured according to the parameters chosen by the ESP teacher such as grade levels.
12) Transfer data to information. This stage consists on the design of the learning activity headings, the type of activity. The corpus data have been transformed into information which is exploited for pedagogical purposes.

4.3. MARNAV Corpus-based instructional material.

A specialised corpus on Maritime Navigation was compiled ad hoc. It consists of 131 texts and 11558 tokens from specialisation level 1. It can be described as a specialised written monolingual English corpus, modular and untagged. The corpus will be hosted by a web application specifically developed to transfer corpus analysis to the context of specialised language learning and teaching. The tool is tailored to specific needs in the LSP setting of the domain of maritime navigation, seafaring, naval architecture, engineering and applied sciences but it can be expanded to any other domain. However, it has not been fully implemented yet. Although the automatically generated activities cannot be displayed at present, figures 2, 3 and 4 below are sample models of how the learning activities are envisaged and can be projected onto the programme. Part 1 illustrates the pre-warm corpus-based data driven learning activity aimed at learner's observation and awareness of the term behaviour, followed by another task that requires the learner's interaction with the processing tool to obtain results.

1. Part 1 (DDL):
    2. Upload the file named "Marine navigation", click on the "word list" tab and observe the most frequent words in the document. Then, click on concordance tab and write the word "ship" in the search box. You will see a list of sentences containing the word ship.

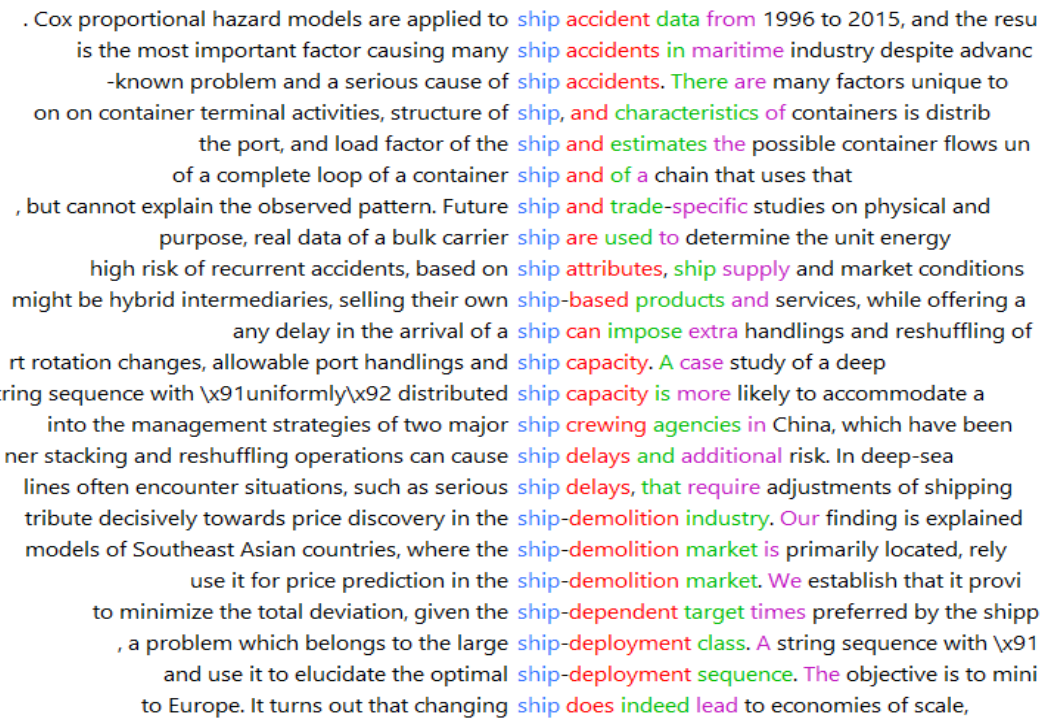

Figure 2. The resulting concordance search for the word 'ship' using AntConc (2017)

    a. Highlighted in red colour you will see the words that usually accompany 'ship'. Click on the "accident" hit in the second line and read the full paragraph. Then, work with your class mate and explain what the paragraph is about.

According to the latest BIMCO and Drewry reports, there is a global shortage of officers for the worlds\x92 merchant fleet. some of the maritime education and training challenges facing these officers in accessing global labour markets. The paper important actor in shaping global labour markets. Using a qualitative approach, interviews were conducted with 10 key inf Interview data was analysed and coded for themes using NVivo qualitative data analysis software (QSR International Pty Lt Clarke\x92s six-step method of thematic analysis. This was combined with a review of labour market statistics to demonstr berth availability and the lack of South African ship ownership. The solutions adopted by the state includes a limited adopti Australia, Singapore, Taiwan and Nigeria. The article contributes to filling the gap in empirical-based maritime studies that

Human error is the most important factor causing many ship accidents in maritime industry despite advanced technology serious cause of ship accidents. There are many factors unique to the marine environment raising the potential for fatigue fatigue to be a root cause of accident, it is important to devise methods to detect and quantify the fatigue and mental symp measuring fatigue level and \x91Symptom Checklist 90- Revised\x92 (SCL-90-R) for detecting the severity of mental symp Social Sciences) software. According to the results of PFS analysis, a slight degree of fatigue is detected in all sub-dimensio mental symptoms perceived by seafarers is not generally highly detected. In conclusion, the purpose of this study is to dete symptoms among seafarers caused by working conditions on-board.

Figure 3. Sample text extracted from the corpus showing occurrences for the word 'ship'

3. Part 2 (DDL):

   Answer the following questions:

   - How many types of turbines can you locate in the text? how many types of generators?
   - Match the following terms with the corresponding abbreviation that appears in the text:
     a. Technique for Order Preference by Similarity to Ideal Solution
     b. wind turbine generators
     c. fuzzy analytic hierarchy process
     d. power system simulator for engineering

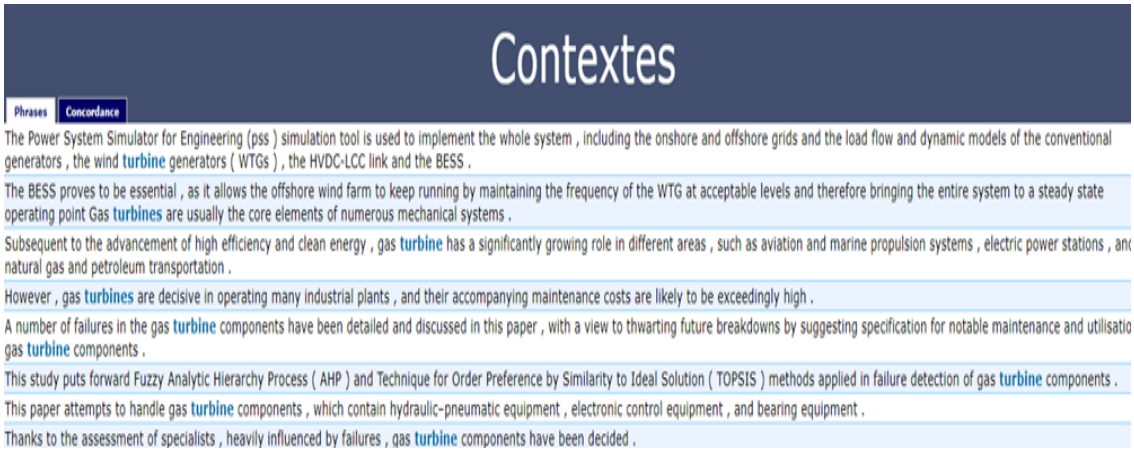

Figure 4. The resulting search for the word 'turbine' using *TermStat* context tool

The second part above delineates a basic corpus-based task on terminology learning consisting on locating, identifying and matching abbreviations to terms.

5. CONCLUDING REMARKS

Specialised language corpora contain data to inform ESP instructional material on the key terminological units whose frequency and co-occurrence deserve special attention as their specificity degree is clearly identifiable in a given (sub) domain. The dataset yielded by a corpus provides the ESP

teacher with sufficient input to develop purposeful corpus-based tasks. Corpus exploitation for pedagogical ends is empirically grounded and, therefore, entirely appropriate to be performed within a sustainable greater pedagogical framework such as Data Driven Learning (Johns 1991), as it has been demonstrated by previous research tradition. To our knowledge, few studies have focused on providing guidelines to build up corpus particularly tailored to the ESP pedagogical needs and backed up by a detailed workflow from the envisage of the corpus design to its final learning/teaching material, that is from grasping corpus data to its actual exploitation. This paper provides the ESP teacher/practitioner with guidelines to control the quality of the corpus compilation and a workflow of actions from the documentary stage to corpus processing and data extraction, and intercepted by a series of actions before the transfer of corpus data to information takes place. Our main contribution can be summarised as the importance of establishing a general quality protocol in which a specialised corpus can be developed. What seems clear from this study is that despite Corpus Linguistics gives us the keys to obtain quantitative and qualitative data empirically demonstrated, pedagogical applications also involves the frame of a methodological approach. In this vein, corpus-informed Data Driven Learning modes entail both intentional and accidental discovery of terminological information. Among the main advantages it is worth mentioning the great deal of information that corpus can provide.

Even if the designer focuses attention on terms, new data are continuously retrieved through co-occurrence patterns so that lexico-grammar, semantic and extralinguistic information can be drawn. The ESP learner can benefit from the visualisation of the term environment enhancing memory performance through contextual information. Written samples of corpora also help students focus on patterns for improving writing skills. The learner's sustained interest and strong motivation are also fostered through the use of technological tools and hands-on training. On the other hand, this learner-centred approach may also contribute to enhance the learner's language awareness in the professional field. As regards the ESP teachers, corpus-informed DDL contributes to teacher development, enhancing the language awareness and research skills. It also provides them with self-assurance and autonomy as material developers and controllers of a pedagogically oriented process and of its final product.

The main drawbacks that should be mentioned are the lack of availability of resources, the restricted access to documentary sources, the dependency on the subject-matter expert to validate terms and elucidate prototypical genres and text types of the (sub) domain, the necessary skills to interpret the quantitative data provided by the computational tools and the time constraints.

In our view, it could be an interesting topic for future research to detect and analyse the errors that emerged during the learner's performance and review the instructional material looking for inconsistencies as well. On the other hand, expectations on the full implementation of the programme to automatically generate corpus-informed learning activities will allow piloting studies and practical cases for further discussions. Future investigations that stem from experimental uses of our proposal are necessary to validate and confirm these initial findings.

ACKNOWLEDGEMENT

The research reported in this work has been funded by 2016-2017 Post-doctoral Research Grant of the Universidad de Cádiz, Spain (grant number EST2017-138).

REFERENCES

Anthony, L. 2017. *AntConc* (Version 3.5.2) [Computer Software]. Tokyo, Japan: Waseda University. Available at http://www.laurenceanthony.net/software

Atkins, S., Clear, J., and Ostler, N. 1991. «Corpus Design Criteria». *Literary & Linguistic Computing*, 7/1, 1-16.

Bernardini, S. 2000. *Competence, Capacity, Corpora. A Study in Corpora Aided Language Learning.* Bolonia: Clueb.

Biber, D., 1993. «Representativeness in Corpus Design». *Literary and Linguistic Computing*, 8/4, 243-257.

Biber, D., Conrad, S. and Reppen, R. 1998. *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Boulton, A. and Pérez-Paredes, P. 2014. «Researching uses of corpora for language teaching and learning». Editorial. *ReCALL*, 26/2, 121-127. <hal-00943861>.

Bowker, L. and Pearson, J. 2002. *Working with Specialised Language: a Practical Guide to Using Corpora.* London: Routledge.

Cabré Castellví, M. T. 1999. *La Terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos.* Barcelona: Institut Universitari de Lingüística Aplicada. Barcelona: Universitat Pompeu Fabra.

Campoy, M., M. Gea-valor and B. Belles-Fortuno. 2010. *Corpus-based Approaches to English Language Teaching.* London: Continuum.

Chirobocea, O. 2017. «The Good and the Bad of the Corpus-Based Approach (or Data-Driven Learning) to ESP Teaching». *Mircea cel Batran Naval Academy Scientific Bulletin*, 20/1, 364-371.

Corpas Pastor, G. 2001. «Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada». *Trans*, 5, 155-184.

Corpas Pastor, G. and Seghiri Domínguez, M. 2007. «Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor». *Procesamiento del Lenguaje Natural*, 39, 165-172.

Costa, R. and Silva, R. 2004. «The Verb in the Terminological Collocations. Contribution to the Development of a Morphological Analyser. MorphoComp». *Conference Proceedings of the Fourth International Conference on Language Resources and Evaluation* (Lisbon, Portugal, 26-28 May 2004). Available at http://www.lrec-conf.org/proceedings/lrec2004/ [Accessed on 16.07.2016]

Curado-Fuentes, A. 2002. Exploitation and assessment of a Business English Corpus through language learning tasks. *ICAME Journal*, 26, 5-31. Available at http://www.hit.uib.no/icame/ij26/curadofuen.pdf. [Accessed on 12.11.2017].

Flowerdew, L. 2012. *Corpora and Language Education.* New York: Palgrave Macmillan.

Gabrielatos, C. 2005. «Corpora and Language Teaching: Just a Fling or Wedding Bells? ». *Teaching English as a Second Language* - Electronic Journal, 8/4. Available at: http://tesl-ej.org/ej32/a1.html [Accessed on 12.12.2016]

Ji-Yeon, C. 2014. «The Use of General and Specialised Corpora as Reference Sources for Academic English Writing: A Case Study». *ReCall*, 26/2, 243-259.

Johns, T. 1991. «Should you be persuaded: Two examples of data-driven learning». *English Language Research Journal*, 4, 1-16.

Laviosa, S. 1988. «The Corpus-based Approach: A New Paradigm in Translation Studies». *Meta: Journal des Traducteurs/ Meta: Translation Journal*, 43/4, 474-479.

Leech, G. 1997. «Teaching and Language Corpora: A convergence». In A. Wichman, S. Fligelstone, T. McEnery & G. Knowles (Eds.). *Teaching and Language Corpora*. New York: Addison Wesley Longman, 1-23.

Lewis, M. 1997. *Implementing the Lexical Approach: Putting Theory into Practice.* Hove, England: Language Teaching Publication.

Losey-León, M. A. 2015. «Corpus Design and Compilation Process for the Preparation of a Bilingual Glossary (English-Spanish) in the Logistics and Maritime Transport Field: LogisTRANS». *Procedia: Social and Behavioral Sciences*, 173, 293-299.

Luzón Marco, M. J. 2000. «Collocational Framework in Medical Research papers: a Genre-based Study». *English for Specific Purposes*, 19, 63-86.

Marinov, S. 2013. «Training ESP Students in Corpus Use – Challenges of Using Corpus-based Exercises with Students of Non-philological Studies». *Teaching English with Technology*, 13/4, 49-76. Available at http:///www.tewtjournal.org [Accessed on 12.01.2018].

Marzá, N. E. 2014. «A Practical Corpus-based Approach to Teaching English for Tourism». *International Journal of Applied Linguistics and English Literature*, 3/1, 129-136.

Mizumoto, A., Hamatani, S., and Imao, Y. 2017. «Applying the Bundle-Move Connection Approach to the Development of an Online Writing Support Tool for Research Articles». *Language Learning*, 67/4, 885-921. Available at https://onlinelibrary.wiley.com/doi/abs/10.1111/lang.12250 [Accessed on 08.02.2018]

Scott, M., and C. Tribble. 2006. *Textual Patterns: Keyword and Corpus Analysis in Language Education.* Amsterdam: Benjamins.

Seghiri, M. 2017. «Metodología de elaboración de un glosario bilingüe y bidireccional (ingles/español; español/ingles) basado en corpus para la traducción de manuales de instrucciones de televisores». *Babel*, 63/1, 43-64. Available at https://www.jbe-platform.com/content/journals/10.1075/babel.63.1.04seg [Accessed on 16.02.2019]

Sinclair, J. M. 1987. *Collins Cobuild English Language Dictionary.* London: Collins.

Willis, J., Willis, D. 1988. *Collins COBUILD English Course*. London: Collins ELT.

Wu, L.-F. 2014. «Motivating College Student's Learning English for Specific Purposes Courses through Corpus Building». *English Language Teaching*, 7/6, 120-127.