Using Data-Driven Learning Methods in Telecommunication English Teaching

Camino Rea Rizzo¹, María José Marín Pérez²
Universidad Politécnica de Cartagena (España)¹, Universidad de Murcia (España)²

Abstract

The use of linguistic corpora in language teaching has spread considerably in the last twenty-five years thanks to the pioneer work by Johns [1], who coined the term data-driven learning (DDL, henceforth); Sinclair [2], who developed the concept further on; or Boulton [3], amongst others. DDL teaching methods promote language study based on the observation of concordances, that is, examples of the authentic use of keywords in context (KWC), which are retrieved from a linguistic corpus by running software programs specifically designed to that end, such as Wordsmith [4].

According to the literature on the subject, there exist arguments for and against the use of corpora in language teaching and there has been a fairly small number of pedagogical experiments in English for specific purposes (ESP) [3], particularly in the field of telecommunication English. This work suggests two activities for teaching terminology within this area applying DDL-based methodology, together with a pedagogical experimental model for its future implementation. Such activities are not intended to substitute any other teaching material like course books; they are rather envisaged as a supplement to language exposure and/or reinforcement of terminology. The language samples of this specialized variety are stored in the Telecommunication English Corpus (TEC) [5], designed and compiled ad hoc for language research owing to the scarcity of technical corpora available.

1. Introduction

The use of linguistic corpora in language teaching has spread considerably in the last twenty-five years thanks to the pioneer work by Johns [1], who coined the term data-driven learning (DDL, henceforth); Sinclair [2], who developed the concept further on; or Boulton [3], amongst others. DDL teaching methods promote language study based on the observation of concordances, that is, examples of the authentic use of keywords in context, which are retrieved from a linguistic corpus by running software programs specifically designed to that end, such as Wordsmith [4].

The adequacy of applying DDL methodologies in ESP has been also supported on the basis that students can access authentic samples of the language used by the professional discourse community [3, 5, 6]. Römer [7] reinforces such argument by introducing a series of DDL studies which “demonstrate that corpora nicely complement existing reference books and they may provide information which a dictionary or grammar book may not provide”. Likewise, specialised corpora become clearly essential to ESP precisely for their specialisation “the more specific the need, the more difficult it is to develop materials that are financially worthwhile” [8].

However, in spite of the interest raised in the use of specialised corpora in language instruction, their availability in many different varieties is certainly limited. Therefore, the Telecommunication English Corpus (TEC) [9] was designed and compiled in a modality which allows full processing and the extraction of the information required for particular research aims, such as the preparation of activities for teaching technical terminology, among other possible uses.

TEC is a fairly representative sample of the professional and academic written English of Telecommunication Engineering which amounts to 5.5 million words. The samples originate from real communication acts where at least one user of the language is an expert or professional. Therefore, a considerable variety of typical texts of the discourse community are included, such as research papers, datasheets, books, reports, news, etc. All of them are classified into eight sections depending on the text origin: magazines, books, web, research papers, abstracts, brochures, advertising and technology news.

Concerning topic representativeness within the realm of telecommunications, the curricula of two university degrees were taken as reference: Telecommunication Engineering and Telematics Engineering at the UPCT¹. Every area of knowledge which the curricula consist of meant a thematic line to gather samples of the language. Subsequently, every area of knowledge is constituted by a number of content subjects which narrow down the scope of the topic search. As a result, the corpus

¹ Universidad Politécnica de Cartagena, Spain.
is structured into eight sections comprising the seven main areas of knowledge (Electronics; Computing Architecture; Telematics; Communication and Signal; Materials Science; and Business Management), plus the specialisations in Telecommunication Networks and Systems, and Telecommunication Planning and Management.

2. General vocabulary versus specialized vocabulary: term extraction lists for in-class teaching
The attempts to generate lists of general vocabulary for language learning go back to the 18th century, although their reliability is questionable for several reasons such as the source of texts, the defining criteria, or even the inclusion of words with different senses [10]. Drawing on the work of Thorndike and Lorge [11], acknowledged as the first authors who consider polysemy when organizing their lists, West [12] provides a list which includes the most frequent 2,000 word families in English: the General Service list of English Words. Later, Coxhead’s work [12] resumes this activity and contributes with the Academic Word List: the most frequent 570 word families in academic English.

The existence of lexical inventories is a basic source of information as they allow to sequence vocabulary in language teaching. The organization of general vocabulary with respect to frequency facilitates the development of tests to measure the level of vocabulary knowledge, and therefore, to determine the amount of information learners are able to understand in the second language [14].

With regard to telecommunication English, there are not standard lists which the ESP teacher can rely on to develop vocabulary activities or tests, so this compelling need can be satisfied by analysing a specific corpus and retrieving a lexical repertoire. Therefore, the telecommunication engineering world list (TEWL) was generated by Rea [9] under a corpus-comparison approach, that is, a smaller specific corpus (TEC) is compared to a larger general language corpus (LACELL) which establishes the norm. This approach allows to apply statistical tests to quantify occurrence probability and representativeness of lexical units. TEWL holds 402 specialised families plus 1,017 individual specialised forms which are all found within the range of the 1,000 most statistically significant word families. Additionally, the words in TEWL comply with the quantitative conditions which determine a technical term [15]: a lexical unit must be at least 50 times more frequent in the specialised register than in general language. The list includes the most salient, central and typical specialised lexical units in telecommunications, and corresponds both to words whose use is restricted to the subject-domain, and those which activate a specialised meaning in the discipline even though they may be also used in other fields or in general language.

3. Activities proposal
In keeping with Harwood’s [6] and Boulton’s [8] assertion that corpora properly complement other teaching materials for ESP, two TEC-based activities are designed so that they could enrich textbooks for telecommunication English like English for ICT studies [16]. This book is suitable for B2 level students who, according to the CEFR for languages5, should be able to understand the main ideas of complex texts on both concrete and abstract topics, including technical discussions in their field of specialisation. The proposed activities are aimed at students in the third year of the degree in Telecommunication/Telematics Engineering at UPCT whose curricula include technical English as a compulsory subject.

For the implementation of the activities and advancing the possible outcome, a random selection of students with a minimum level of English (B1) was required. Therefore, a placement test is necessary to discard lower level students and, in turn, classify them depending on their language command in order to observe the potential results of the activities and compare their progress. Subsequently, a test intended to measure the knowledge on the contents dealt with in the activities was designed to be administered before and after doing the suggested activities. This would allow to compare pre and post-test results and gauge the effect of the activities.

As for the integration of activities within the syllabus of the course, they were designed to meet the contents of a particular unit at convenience. Thus, Activity 1 could be performed within the practicals related to unit 1 as they deal with analogue and digital transmissions, where the terms transmit and configure are widely employed. Activity 2 would be scheduled towards the end of the semester, when students have already gained greater knowledge of the field, since they need at least some working knowledge of the different subdomains of telecommunications.

Different levels of the language can be explored through the use of corpora in class. The activities suggested below focus particularly on the morphological, syntactic and semantic levels.

---

5 http://www.um.es/grupolacell/proyectos/
3 http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf
3.1 Activity 1

Students are asked to form the word families [17] of a number of terms extracted from TEWL. Students should therefore do a search on the corpus of possible derivatives by adding prefixes and suffixes to confirm their intuition. One of the advantages of this kind of activity is to make students think consciously about the mechanisms of word formation, encouraging the subsequent application of the rules inferred in similar cases. Such derivation exercises are common in text books, but they often lack a context that makes the terms belonging to each family more meaningful. It is precisely at this point that corpus-based activities provide an added value, since the series of concordances retrieved offer the specific use of the derivational terms in their context.

Once the derivatives are identified, students do a fill-in-the-gaps exercise in which, using gapped concordance lines selected and filtered by the teacher, students complete the sentences with the appropriate terms. This activity would act as feedback confirming what they have previously learned inductively from the observation of real language samples obtained from the corpus. Figures 1 to 4 illustrate some concordance lines extracted for the derivatives of the verbs *transmit* and *configure*.

---

**Figure 1. Transmitance concordance lines.**

---

**Figure 2. Retransmit concordance lines.**

---

**Figure 3. Configuration concordance lines.**

---

**Figure 4. Reconfigurability concordance lines.**
3.2 Activity 2

The second activity explores the syntactic and semantic levels of the language and introduces the concept of collocation, that is, a sequence of words that co-occur more often than would be expected by chance. As stated by Sinclair [18]: “the [statistically significant] occurrence of two or more words within a short space from each other in a text.” Wordsmith allows to retrieve both the collocates and frequency counts of a search word. Such lexical behaviour contributes to construct meaning and specialisation in a particular domain. Therefore, students are asked to observe the collocates and patterns of a term in several subdomains of telecommunications covered in TEC. After sorting and selecting the most frequent and significant collocates of the term, i.e. network (figure 5), students should be able to identify the subdomain in which the term is being used. They can also resort to the source text to expand the context of usage if necessary. Figure 5 illustrates the collocational pattern of network in signal processing, figure 6 in electronics and figure 7 in business.

Figure 5. Network in signal processing.

Figure 6. Network in electronics.
4. Conclusion
This study has presented a proposal to implement data-driven methodology for teaching telecommunication English. Due to the scarcity and restricted availability of existing specific corpora, a corpus of telecommunication English is designed ad hoc as a database. Moreover, owing to the lack of specialised telecommunication term inventories and the fact that they can be useful for, among others, the sequencing of activities to learn specialised vocabulary, Section 3 suggests two activities for teaching telecommunication terminology based on corpus materials obtained from TEC. Such activities correspond to a pilot experience which ought to be expanded along the course by adding a greater number of corpus-based activities. As further research, it would be highly desirable to implement these activities coupled with a pre- and post-test to try and measure the success of the data-driven methodology in the long term. It would also be advisable to compare DDL methods with more traditional ones and decide on the relevance and suitability of resorting to language corpora as a complement to already existing materials with a more prescriptive character.

References

