International Conference ICT for Language Learning

Using Data-Driven Learning Methods in Telecommunication English Teaching

Camino Rea Rizzo[1], María José Marín Pérez[2]
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 UPCT[1]. 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[2]) 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 languages[3], 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. Activity2 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.

[2] 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*.

Concord
File Edit View Compute Settings Windows Help

N	Concordance
1	structure. The absorbance in Ith layer Al(() can be written as:(7.15) and transmittance through the first I layers, Tl((), can be expressed a
2	of the medium (air or glass) above the top boundary of the assembly. Total transmittance T(() can be expressed as: (7.12) where (j is the rat
3	depends on the wavelength (of the incident radiation, the reflectance, transmittance and absorbance calculated in this way also depen
4	graph of Fig. C-26b, the dependence of X3 onX2 is described by he branch transmittance Gc and that of X2 on X3 by the branch transmittan
5	are represented by nodes. The directed arrow is the branch whose branch transmittance G defines the functional relationship X2 = GXl. The
6	(Wm-2nm-1), the normalized integrated reflectance , normalized integrated transmittance through l layer, and normalized integrated absorba
7	Example 1.6.1: Evanescent wave B. Intensity, Reflectance and Transmittance Example 1.6.2: Reflection of light from a less den

Figure 1. *Transmitance* concordance lines.

Concord
File Edit View Compute Settings Windows Help

N	Concordance
1	MH to send duplicated acks that trigger the fast retransmit mechanism of TCP. In this case, the
2	mechanism of TCP. In this case, the fast retransmit of TCP unnecessarily retransmits a
3	before the cwnd is large enough to allow for fast retransmit, will result in a timeout at the server.
4	distracting noise in the background, people will retransmit themselves, usually in the manner of,
5	packets are flowing to cause a TCP fast retransmit.) An operation (involving a single
6	algorithm performs better. Using a per-chunk retransmit bit instead of the timestamp chunk
7	. Here, the protocol may have the sender retransmit a certain piece of data missed by the

Figure 2. *Retransmit* concordance lines.

Concord
File Edit View Compute Settings Windows Help

N	Concordance
1	modern FPGAs offer features like partial reconfiguration, these kinds of hardware
2	software) that's updated with FPGAs. This reconfiguration of the hardware is a serious
3	capabilities raise the prospect of on-the-fly reconfiguration in an operational system
4	remote, lights-out access to permit dynamic reconfiguration, expansion and upgrades.
5	mechanism for incorporating dynamic reconfiguration capabilities into arbitrary
6	of this approach are the dependence of reconfiguration on the dimensions of the
7	the desired target array, are presented. Two reconfiguration algorithms are presented.

Figure 3. *Configuration* concordance lines.

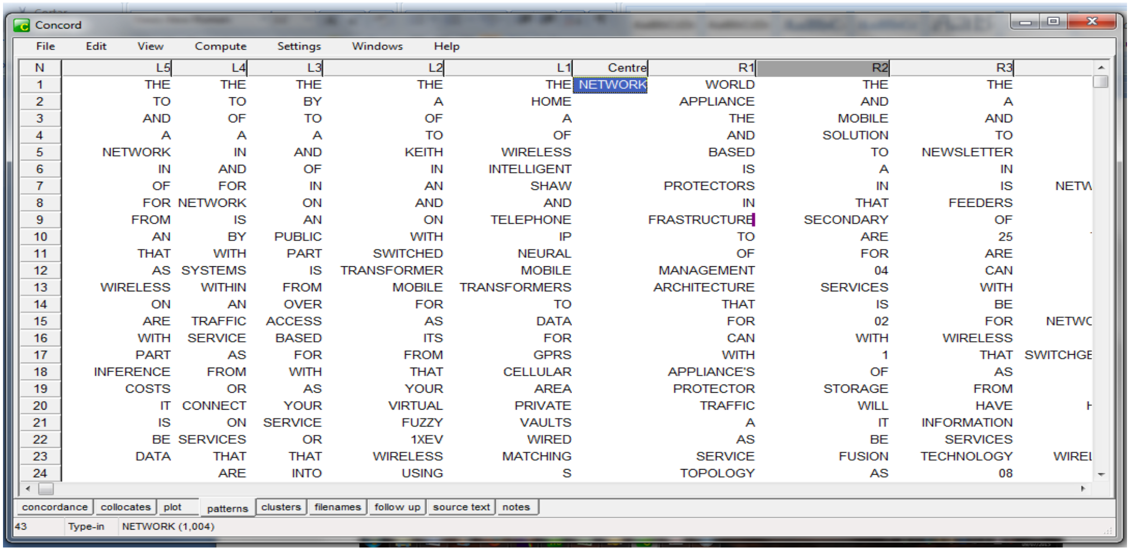
Concord
File Edit View Compute Settings Windows Help

N	Concordance
1	include 2-, 4- or 8-channel or larger configurations. "Reconfigurability is a key requirement of
2	. Author Keywords: Reconfiguration; hexagonal array; reconfigurability; fault tolerance. Direct connect device
3	are based on the need of programmability or dynamic reconfigurability in order to extend the life-time of a
4	are based on the need of programmability or dynamic reconfigurability in order to extend the life-time of a
5	, seamless integration and interoperability, rapid re-reconfigurability and flexibility for future growth.
6	during inter-processor communication. This reconfigurability of PCA may help satisfy varying
7	include 2-, 4- or 8-channel or larger configurations. "Reconfigurability is a key requirement of

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.

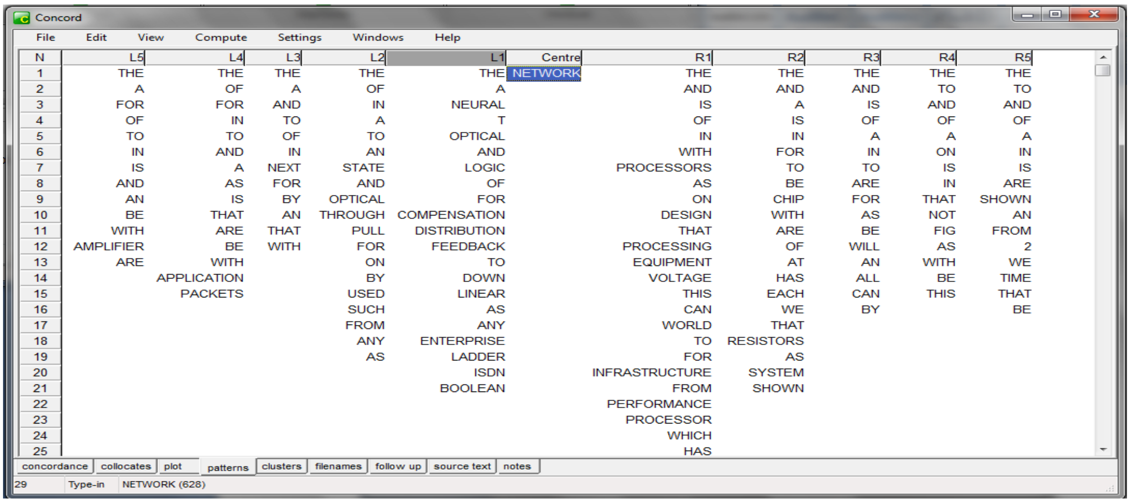


Concord
File Edit View Compute Settings Windows Help

N	L5	L4	L3	L2	L1	Centre	R1	R2	R3	
1	THE	THE	THE	THE	THE	NETWORK	WORLD	THE	THE	
2	TO	TO	BY	A	HOME		APPLIANCE	AND	A	
3	AND	OF	TO	OF	A		THE	MOBILE	AND	
4	A	A	A	TO	OF		AND	SOLUTION	TO	
5	NETWORK	IN	AND	KEITH	WIRELESS		BASED	TO	NEWSLETTER	
6	IN	AND	OF	IN	INTELLIGENT		IS	A	IN	
7	OF	FOR	IN	AN	SHAW		PROTECTORS	IN	IS	NETW
8	FOR	NETWORK	ON	AND	AND		IN	THAT	FEEDERS	
9	FROM	IS	AN	ON	TELEPHONE		FRASTRUCTURE	SECONDARY	OF	
10	AN	BY	PUBLIC	WITH	IP		TO	ARE	25	
11	THAT	WITH	PART	SWITCHED	NEURAL		OF	FOR	ARE	
12	AS	SYSTEMS	IS	TRANSFORMER	MOBILE		MANAGEMENT	04	CAN	
13	WIRELESS	WITHIN	FROM	MOBILE	TRANSFORMERS		ARCHITECTURE	SERVICES	WITH	
14	ON	AN	OVER	FOR	TO		THAT	IS	BE	
15	ARE	TRAFFIC	ACCESS	AS	DATA		FOR	02	FOR	NETWC
16	WITH	SERVICE	BASED	ITS	FOR		CAN	WITH	WIRELESS	
17	PART	AS	FOR	FROM	GPRS		WITH	1	THAT	SWITCHGE
18	INFERENCE	FROM	WITH	THAT	CELLULAR		APPLIANCE'S	OF	AS	
19	COSTS	OR	AS	YOUR	AREA		PROTECTOR	STORAGE	FROM	
20	IT	CONNECT	YOUR	VIRTUAL	PRIVATE		TRAFFIC	WILL	HAVE	H
21	IS	ON	SERVICE	FUZZY	VAULTS		A	IT	INFORMATION	
22	BE	SERVICES	OR	1XEV	WIRED		AS	BE	SERVICES	
23	DATA	THAT	THAT	WIRELESS	MATCHING		SERVICE	FUSION	TECHNOLOGY	WIREL
24		ARE	INTO	USING	S		TOPOLOGY	AS	08	

concordance | collocates | plot | patterns | clusters | filenames | follow up | source text | notes
43 Type-in NETWORK (1,004)

Figure 5. *Network* in signal processing.

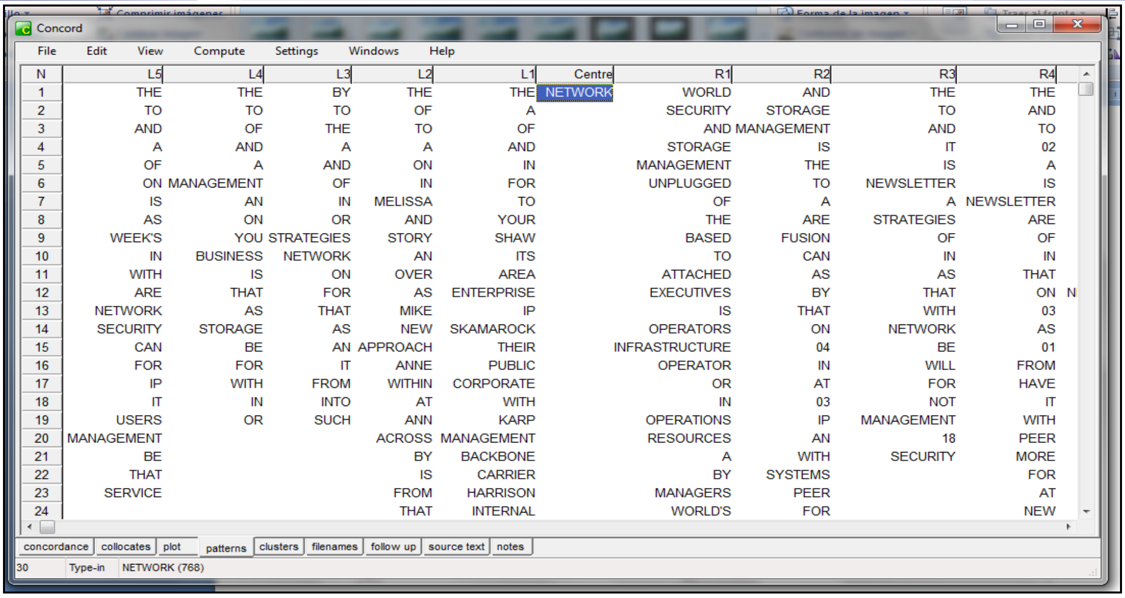


Concord
File Edit View Compute Settings Windows Help

N	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	THE	THE	THE	THE	THE	NETWORK	THE	THE	THE	THE	THE
2	A	OF	A	OF	A		AND	AND	AND	TO	TO
3	FOR	FOR	AND	IN	NEURAL		IS	A	IS	AND	AND
4	OF	IN	TO	A	T		OF	IS	OF	OF	OF
5	TO	TO	OF	TO	OPTICAL		IN	IN	A	A	A
6	IN	AND	IN	AN	AND		WITH	FOR	IN	ON	IN
7	IS	A	NEXT	STATE	LOGIC		PROCESSORS	TO	TO	IS	IS
8	AND	AS	FOR	AND	OF		AS	BE	ARE	IN	ARE
9	AN	IS	BY	OPTICAL	FOR		ON	CHIP	FOR	THAT	SHOWN
10	BE	THAT	AN	THROUGH	COMPENSATION		DESIGN	WITH	AS	NOT	AN
11	WITH	ARE	THAT	PULL	DISTRIBUTION		THAT	ARE	BE	FIG	FROM
12	AMPLIFIER	BE	WITH	FOR	FEEDBACK		PROCESSING	OF	WILL	AS	2
13	ARE	WITH		ON	TO		EQUIPMENT	AT	AN	WITH	WE
14		APPLICATION		BY	DOWN		VOLTAGE	HAS	ALL	BE	TIME
15		PACKETS		USED	LINEAR		THIS	EACH	CAN	THIS	THAT
16				SUCH	AS		CAN	WE	BY		BE
17				FROM	ANY		WORLD	THAT			
18				ANY	ENTERPRISE		TO	RESISTORS			
19				AS	LADDER		FOR	AS			
20					ISDN		INFRASTRUCTURE	SYSTEM			
21					BOOLEAN		FROM	SHOWN			
22							PERFORMANCE				
23							PROCESSOR				
24							WHICH				
25							HAS				

concordance | collocates | plot | patterns | clusters | filenames | follow up | source text | notes
29 Type-in NETWORK (628)

Figure 6. *Network* in electronics.



N	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4
1	THE	THE	BY	THE	THE	NETWORK	WORLD	AND	THE	THE
2	TO	TO	TO	OF	A		SECURITY	STORAGE	TO	AND
3	AND	OF	THE	TO	OF		AND	MANAGEMENT	AND	TO
4	A	AND	A	A	AND		STORAGE	IS	IT	02
5	OF	A	AND	ON	IN		MANAGEMENT	THE	IS	A
6	ON	MANAGEMENT	OF	IN	FOR		UNPLUGGED	TO	NEWSLETTER	IS
7	IS	AN	IN	MELISSA	TO		OF	A	A	NEWSLETTER
8	AS	ON	OR	AND	YOUR		THE	ARE	STRATEGIES	ARE
9	WEEK'S	YOU	STRATEGIES	STORY	SHAW		BASED	FUSION	OF	OF
10	IN	BUSINESS	NETWORK	AN	ITS		TO	CAN	IN	IN
11	WITH	IS	ON	OVER	AREA		ATTACHED	AS	AS	THAT
12	ARE	THAT	FOR	AS	ENTERPRISE		EXECUTIVES	BY	THAT	ON
13	NETWORK	AS	THAT	MIKE	IP		IS	THAT	WITH	03
14	SECURITY	STORAGE	AS	NEW	SKAMAROCK		OPERATORS	ON	NETWORK	AS
15	CAN	BE	AN	APPROACH	THEIR		INFRASTRUCTURE	04	BE	01
16	FOR	FOR	IT	ANNE	PUBLIC		OPERATOR	IN	WILL	FROM
17	IP	WITH	FROM	WITHIN	CORPORATE		OR	AT	FOR	HAVE
18	IT	IN	INTO	AT	WITH		IN	03	NOT	IT
19	USERS	OR	SUCH	ANN	KARP		OPERATIONS	IP	MANAGEMENT	WITH
20	MANAGEMENT			ACROSS	MANAGEMENT		RESOURCES	AN	18	PEER
21	BE			BY	BACKBONE		A	WITH	SECURITY	MORE
22	THAT			IS	CARRIER		BY	SYSTEMS		FOR
23	SERVICE			FROM	HARRISON		MANAGERS	PEER		AT
24				THAT	INTERNAL		WORLD'S	FOR		NEW

Figure 7. *Network* in business.

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

[1] Johns, T. (1986). Microconcord:A language-learner's research tool. *System*, 14 (2), 151–162.
[2] Sinclair, J. (2003). *Reading Concordances:An Introduction*. London:Longman.
[3] Boulton, A. (2010). Data-driven learning:Taking the computer out of the equation. *Language Learning*, 60, 534-572.
[4] Scott, M. (2008). *WordSmith Tools* [computer software]. Liverpool:Lexical Analysis Software.
[5] McEnery, T., Wilson, A. (1996). *Corpus Linguistics*. Edinburgh:Edinburgh University Press.
[6] Harwood, N. (2005). What do we want EAP teaching materials for? *Journal of English for Academic Purposes*, 4, 149-161.
[7] Römer, U. (2008). Corpora and Language Teaching. In Lüdeling, A. and Kyto, M. (eds). *Corpus Linguistics. An International Handbook, Volume I* (112-130). Berlin:Mouton de Gruyter.
[8] Boulton, A. (2012). Corpus consultation for ESP. A review of empirical research. In Boulton, A., Carter-Thomas, S., Rowley-Jolivet, E. (eds.) *Corpus-Informed Research and Learning in ESP.* Issues and Applications (261-292). John Benjamins Publishing Company.
[9] Rea, C. (2008). *El inglés de las telecomunicaciones: estudio léxico basado en un corpus específico*.http://www.tesisenred.net/TDR-0611109-134048/index_cs.html
[10] Sánchez, A. (2000). Language Teaching before and after 'Digitilized Corpora'. *Cuadernos de Filología Inglesa*, 9.1. Murcia:Servicio de Publicaciones.
[11] Thorndike, E.L., Lorge, I. (1944). *The teacher's Word Book of 30,000 Words*. Teachers' College:Columbia University.
[12] West, M. (1953). *A General Service List of English Words*. London:Longman.
[13] Coxhead, A. (2000). A New Academic Word List. *TESOL, Quarterly*, 34 (2), 213-238.
[14] Nation, I.S.P., Beglar, D. (2007). A vocabulary size test. *Language Teacher*, 31 (7), 9-13.

[15] Chung, T.M. (2003). A corpus comparison approach for terminology extraction. *Terminology*, 9 (2), 221-246.
[16] Fitzgerald, P. et al. (2011). *English for ICT Studies*. Garnet.
[17] Bauer, L., Nation, I.S.P. (1993). Word Families. *International Journal of Lexicography*, 6 (4), 253-279.
[18] Sinclair, J. (1991). *Corpus, Concordance and Collocation*. Oxford:Oxford University Press.