Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 459 – 464

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Designing, Describing and Compiling a Corpus of English for Architecture

Begoña Soneira Beloso*

Facultad de Filología Inglesa, Universidad de Santiago de Compostela, Av. Castelao S/N, Santiago de Compostela, 15782

ABSTRACT

This full paper presents the **CADCE** (Corpus of Architecture Discourse in Contemporary English), a collection of approximately 500.000 words of written language from a range of different sources designed to represent the language of Architecture in contemporary English in order to study the lexis of this particular field. The CADCE is monolingual and it is not annotated; it includes texts from North-American, British, Irish, Canadian and Australian publications. It is a synchronic corpus since it gathers recent texts published from 2007-2008. It is a specific-purpose corpus limited to a particular subject, namely Architecture, a discipline that comprises many other related subareas: construction, urbanism, landscape architecture, building materials, green architecture, interior design, etc. The design principles that favoured the creation of this corpus are *representativeness* (size, topic, sources, level of technicality), *contemporariness* (current, authentic, up-to-date publications) and *accessibility* (online, free-accessed, computerized texts). Thanks to the design of this research tool the main aspects involved in the lexical profile of Architecture English could be described and analyzed.

Keywords: Corpus Studies; English for Specific Purposes; Applied Linguistics; Lexicography.

* Corresponding author. Tel.: +34686850854
E-mail address: begona.soneira@usc.es

doi:10.1016/j.sbspro.2015.07.466

1. Introduction

The research study leading to the creation of this corpus aimed at analyzing the English lexis of Architecture Discourse (AD from now on) in order to describe and characterize it. As pointed out by Biber et al (1998), the use of a corpus is the best way to analyze registers, dialects, styles, etc. (here a specialized language) in terms of their linguistic patterns. According to Conrad (2002:77), corpus design (size of the corpus, types of texts included, number of texts, the sampling procedure, etc.) is crucial in order to achieve reliable results. Out of the many different types of corpora, it seemed logical to design what has been called "special purpose corpus" (Pearson, 1998:46) since general reference corpora are rather meant to be representative of all relevant varieties of the language and not of a technical or professional field in particular. Since at the time there was no existing corpus with the features explained above in the field of Architecture, a special purpose corpus was to be compiled: in order to do so a preliminary pilot corpus was required to provide the general guidelines to be applied in the final one.

2. Pilot corpus

A model corpus would throw some light on the most salient phenomena, what was to be looked for, etc. and it would ultimately prove the validity of the corpus as a tool for this particular analysis. Following this a priori drive, a preliminary collection of texts was withdrawn from online publications on Architecture, a list provided by the Colexio de Arquitectos de Galicia. This initial search was based on electronic resources since it was assumed from the beginning that this instrument would be accurate in capturing up-to-date materials and it would be expected to be closer to the real linguistic environment of Architecture professionals whose work is highly computerized and technology-dependent. The outcome was a corpus of 200,000 words. A qualitative analysis of the lexis contributed to generating many linguistic insights on the printouts; some remarkable lexical phenomena were acknowledged, such as morphological neologisms (suburbanization, streetscape, beautification, eco-topian, starchitects, spec-builder, ultra-modish, urban-minded, etc.), terms coined by the use of metaphors (rocket -describing building shapes-, empty gesture -portraying a construction-, etc.), conversions (to freelance, to interface, to trim, to green, etc.), loanwords (protégé, repertoire, virtuoso, inferno, bravura moment, aficionado, etc.), Latinisms (terra incognita, per aspera ad astra, exedra, etc.), use of compounding to coin new technical terms (high-speed fibe-optic data lines, pitched-roofed single-family house, etc.), hybrids (laisser-faire urbanism, architecture parlante, capuccino urbanism, etc.), personifications ("urban surfaces that tell stories"), and other processes.

This pilot search confirmed some of the intuitions and expectations on the wide variety and creativity shown in the lexis of AD and the interest of characterizing this professional language together with the need of establishing clear-cut and documented parameters that would guarantee corpus validity and reliability.

3. Criteria for corpus design

After the pilot corpus, a definitive one was designed following three main features, namely representativeness, contemporariness and accessibility.

3.1. Representativeness and contemporariness

The corpus to be compiled had to accomplish "representativeness" and "contemporariness"; in order to represent what it was intended for, the requirement of size had to be seriously considered. The data bank ought to be large enough as to bring about a significant amount of tokens to be examined and also to allow for generalizations to be derived from it; at the same time, however, the body of data could not be extremely large as to prevent it from being analyzed qualitatively. A final figure of five hundred thousand words was acknowledged as suitable for the expected research needs (a "small corpus" format according to the literature). The general topic of the texts to be collected was Architecture, a subject matter that comprises many other related subareas (construction, urbanism, landscape architecture, building materials, green architecture, interior design, etc.). The level of technicality of the texts was determined by the genre; the texts were meant to be closer to architects' real communication rather than being mere technical inventories or treatises. Another constraint regarding representativeness and contemporariness was related to the sources. The texts taken as input needed to be acknowledged as representative by the very professional group whose discourse was to be analyzed; that is, by architects themselves. Thus the sources chosen are addressed to a readership of practitioners in the domain of Architecture where they are highly renowned. In order to guarantee this requirement, several Associations of Architects in Spain (Colegios de Arquitectos) were approached; these are, to a great extent, familiar with international publications and are the main and most prestigious reference as far as Architecture is concerned in our country. Apart from the reliability, authenticity and currency of the sources, their diversity has to be highlighted as well. Even though national differences were disregarded, it has to be pointed out that the sources of the corpus are quite varied; although most of them are American, there were also publications from England, Ireland, Canada and Australia, a fact that contributes to a certain degree of diversity within this register. I have not included as sources of the corpus those journals coming from non-English-speaking countries which use English as some sort of lingua franca, for they could add an additional dimension to the corpus. However, this does not mean that there might be some texts in the whole bulk of data written by a foreign author although such a circumstance has been considered marginal. Texts chosen are rather recent (2007-2008) and thus they allow us to operate with updated materials, a very important fact when compiling a corpus of disciplinary texts as pointed out by Orna-Montesinos (2012:129) given the evident link of Architecture with technology and design.

3.2. Accessibility

The main concern of this corpus besides being representative is that of being made out of accessible sources (what Orna-Montesinos (2012:129) calls "availability"), that is, accessible to the professionals who work in Architecture and people who are learning and teaching Architecture and thus it will be guaranteed to be real, current, spread and alive, and that it will be up to its community standards. Such accessibility has been achieved by means of the genre and the mode. Within architecture communication, I have chosen to focus on practical or applied texts, a kind of data which lies closer to architects' daily language (as opposed to pure theoretical texts) as pointed out above. The genre chosen was the online press, mostly specialist-non-academic articles on Architecture. I also included other subgenres, namely architectural review (a subgenre describes built projects usually addressed to professionals or people concerned with the architecture field), post-construction assessment, jury citation, interview, exhibition report, architecture book review, editorials, etc. Although this is a case of what Rowley-Jolivet and Campagna (2011:45) and Luzón (2005:28) would call "genre migration" to the Web, since these are journals and magazines accessed online, the texts retain most of the properties possessed in the older medium which become adapted to the new one.

Two other characteristics were taken into account to guarantee accessibility: on the one hand, the material included in the corpus is written, as opposed to spoken, and freely accessed on-line. The nature of the current study does not seem to call for the inclusion of oral discourse. Besides spoken discourse in Architecture is usually linked to communication events which seem to be, at least a priori, less suitable for our purposes than those present in written material. The fact of working with electronic texts contributes as well to the accessibility requirement since they reach a much wider audience. The electronic format also facilitated the texts being computerized. On the other hand, the texts included in the corpus are accessed for free (subscription materials are not entirely accessible outside academic institutions, architecture offices and the like). All journals suggested by the architecture associations had freely available materials; these were the texts I have selected to work with. The communicative setting established for these electronic journals is usually expert-to expert and expert-initiated communication although the fact of being accessed openly online makes these materials available to everyone interested in reaching them.

4. Collecting the data

As has been noted above, I have addressed professional architecture associations in Spain in order to obtain a list of international online publications which are acknowledged in the architect's community to be prestigious journals and magazines all over the world. More specifically, I contacted the documentation departments of the directory found in the CSCAE (Consejo Superior de Colegios de Arquitectos de España); out of each journal, a bulk of approximately forty thousand words was extracted resulting in an overall amount of 359.150 words, a figure that accounts for the main part of the corpus. Full texts were retrieved excluding pictures and captions up to a number of 569 texts. All the texts are preceded by correlative numbers and show their website addresses and dates of compilation (most of them range from January to June 2007 except for the last one, which was compiled in December 2008). This was the amount of texts I could gather from those online journals and magazines in a proportioned way (I could retrieve more texts from some of them but not from all). In order to reach the figure of half a million words a secondary group was added to this primary one; the latter is made of publications that are in turn acknowledged by some of the "primary" ones. To be more precise, twelve more architecture websites were included obtaining a final inventory of twenty one journals. This secondary group provided texts that are in general less institutional and targeted for architecture students/practitioners therefore it complements the first set of texts. The second group contributed to an outcome of 146.499 words (about twelve thousand per journal), thus completing an overall figure slightly over half a million words, which is the total amount of the corpus.

5. Results provided by the corpus

The CADCE was designed and compiled as the methodological foundation to reach a comprehensive description of the lexis of AD in English on three different levels: syntagmatic level, semantic level, borrowing level.

The close analysis of the corpus provided highly valuable information on word-forming processes in AD (compounding, borrowing, derivation, metaphor, specialization, acronymy, analogy, etc.) which proved not to be essentially different from those found in general English or in other specialized languages: the crucial distinction would fall on the frequency in their occurrence and the discourse identity to which they conform. Architecture discourse displays a tendency towards using compacting devices such as compounding (channel glass, engineering brick, fiberglass), due to its attractiveness as a lexical tool grounded on properties, such as concision, interpretability, compactness, semantic transparency and high productivity. Indeed, compounding has proved to be extremely productive means to create new words in the realm of Architecture. Compactness is also obtained through the property of recursivity in many compounds. This lexical density is so common that it represents one of the main factors in the building of AD and its specifically technical nature; long strings of words might also trigger difficulties of comprehension and thus may add lexical obscurity to technical texts, especially from the point of view of non-experts. The corpus shows that the most common type of compounding in AD is nominal compounding. Semantically speaking, however, the most important characteristic in AD as far as compounding is concerned is the power of certain compounds to act as models for inductive generalizations by means of which new compounds are created, that is, compound paradigms which become the model for newly created combinations.

Within the scope of prefixation there are several interesting phenomena such as the use of the negative prefix non- in expressions such as non-spatial or non-place lends these derivatives literary and philosophical undertones. There are also neologisms such as deurbanize, superstructure, or transmaterial. Recent neologisms coined by prefixation are ex-urban, re-zoning or pre-planning, which serve as evidence of the extent to which prefixation is a productive device for the coining of new words in AD. The suffix –ing is quite frequent and is present in recent neologisms such as streamlining, landscaping, daylighting and streetscaping. Another suffix which shows a high degree of productivity is –ism (mainly for styles and movements, as in Palladianism), to be found in relatively new terms such as Miesianism (after architect Mies van der Rohe) and Brutalism. The suffix –ation is still productive, leading to pedestrianiazation and gentrification. Certain significant abstract concepts in the discourse of architecture have been created by means of the suffix –ity, as in materiality, linearity and tactility. It is not rare to find relatively recent adjectives derived from a proper name, usually an architect's name, as in Corbusian, Venturian, Miesian and Framptonian.

In general terms, the impact of backformation and conversion is rather low compared to compounding and derivation; recent examples of conversion include duplex, an adjective that became a noun, and interface, a noun that led to the coining of a verb. Examples of backformation include grid, a backformation from gridiron.

Clipping tends to be more common in the language of Architecture. This word-formation process produces interesting items, such as rehab (rehabilitation), prefab (prefabricated) or self-fab (self-fabricated). Other examples of clipping are Neogoth (Neogothic), decon (deconstructionist) and Low-E. The phenomenon of clipping is also attested in highly technical terms such as tarmac (from tar macadam) and polythene (from polyethylene).

Examples of blends are more restricted in number, although some of those that have arisen are noteworthy as technical expressions, such as tensegrity (tensional integrity) and the material called glulam (glue lamination).

Acronyms are very widespread in the corpus, their usefulness a result of their obvious concision, and their opacity increases the level of technicality of the discourse, e.g. LED, CAD and GPS.

Ex-nihilo formations are not numerous but are salient, particularly in the coining of proper names such as Lucite, Freon and Dymaxion. There are others which have been created randomly, such as Nylon, but which have served as the basis or inspiration of others, in this case Mylar or Kevlar.

There are also cases of analogy such as suburban sprawl, generated by drawing an analogy with another expression, as in urban sprawl.

Another device used for the coining of new words that has been analyzed in this study was loanword neology. Most loanwords found in our architecture corpus come from French and Latin, which, even in recent cases, is a tendency that coincides with what happens in general English. The majority of borrowed words are nouns which are in general highly naturalized, that is, fully adapted to the English language. The overall bulk of architecture terms introduced from abroad are not recent and most of these borrowings have been coined in order to satisfy the need for conceptual accuracy, that is, to fill lexical gaps. More narrowly, there are also loanwords that have been introduced for stylistic reasons. This is done as a means of elevating the style of a text or to introduce elements of rhetorical variation. Loanwords may also portray the essence of the local atmosphere in a text, a fact that illustrates the importance of context for Architecture and its actors, that is, the need to be sensitive towards the “genius loci” of a place (its denoting spirit or essence) by reflecting all its features: language, traditions, history, people, etc.

The study of semantic neology in AD shows the importance of the use of metaphors. Many of the metaphors in Architecture are of the scientific type, from fields as diverse as the natural sciences, Anatomy, Biology, Physics, Mathematics, Mechanics, etc. and, as Forty (2000:11) claims, they belong to the modern era, since it was not in fact until modern times that Architecture and Science were theoretically separated. Within Semantic neology there are also several remarkable processes, such as those of specialization or terminologization (e.g. dense, concept, intervention, mediate, brief, etc.); the migration of terms from a specialized language to Architecture has also been attested in the corpus (e.g. Metabolism, green, ersatz, charette, etc.), a fact that highlights the complex nature of AD, its intellectual ambition, and its interdisciplinary nature. We can also notice the “popularization” of architecture terms into general language, subsumed into the general culture, probably the result of Architecture having become of general interest in society.

6. Conclusions

Although there is a varied literature on the description of specialized languages, to my knowledge none has been devoted wholly to the lexis of Architecture. The CADCE is thus an innovation in this specific field from a methodological perspective since at the time it was compiled there was no other corpus with its characteristics. It was developed according to pre-established research aims and combines manual inspection with computerized processing. The corpus was printed out for a preliminary reading, allowing for an overall impression of the data and to the identification of general features and lexical patterns. This opened up other areas of the specialized lexis of Architecture that helped me during the subsequent analysis of the material allowing for a corpus-based study but also has some corpus-driven traits. The use of this self-compiled corpus made it possible to create a detailed lexical description of an ESP variety, English for Architecture.

Although the CADCE has been compiled specifically for a given research topic, the corpus is potentially of use for other studies, not only on the discourse of Architecture from a lexical perspective but from other linguistic approaches.

7. References

Biber, D., Conrad S., and Reppen, R. (1998). *Corpus linguistics. Investigating language structure and use.* Cambridge: Cambridge University Press.

Conrad, S. (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics, 22*, 75 - 95.

Forty, A. (2000). *Words and buildings: a vocabulary of modern architecture*. London: Thames and Hudson.

Pearson, J. (1998). *Terms in context*. Amsterdam: John Benjamins.

Orna-Montesinos, C. (2012). The duality of communicative purposes in the textbook for construction engineering and architecture: A corpus-based study of blurbs. *Atlantis, 34* (2), 125 - 45.

Rowley-Jolivet, E., and Campagna, S. (2011). From print to web 2.0: the changing face of discourse for special purposes. *LSP Journal 2*(2), 44 - 51.

Luzón, M. (2005). Analysis of online product reviews in the field of Computing: A user's perspective. *LSP and Professional Communication 5*(2), 28 - 47.