Keeping up with the digital age: New data sources in research on languages for specific purposes

Amanda Roig-Marín
University of Cambridge
adr41@cam.ac.uk

Abstract

Social media exchanges (for example, via Facebook or Twitter), blogs, and forums, amongst many other electronic genres, have come to be used as relatively *bona fide* testimonies of language use nowadays. The present paper discusses how such language data can be effectively employed in research carried out into languages for specific purposes. Particularly, it describes a project which aims to compile a representative number of texts (in this case, blog posts) in the Spanish fashion panorama in order explore the lexis occurring in the specialised language. It thus details the procedure and criteria followed to create the database, and it tackles some of the challenges which the reader/researcher may encounter in this enterprise.

Keywords: corpora, Internet-based research, blogs, vocabulary in languages for specific purposes.

Resumen

Siguiendo el ritmo de la era digital: nuevas fuentes de datos en la investigación de lenguas para fines específicos

Los intercambios en redes sociales (por ejemplo, a través de Facebook o Twitter), *blogs* y foros, entre otros muchos géneros digitales, en la actualidad han pasado a ser utilizados como testimonios relativamente fidedignos de la lengua. El presente estudio trata cómo tales fuentes pueden ser utilizadas de forma efectiva en la investigación sobre lenguas para fines específicos. En particular, describe un proyecto que pretende compilar un número representativo de textos (en este caso, *posts de blog*) en el panorama de la moda española para explorar el léxico que se da en esta lengua de especialidad. Detalla, pues, el procedimiento y criterios seguidos para crear la base de datos y aborda algunos de los retos que el lector/investigador puede encontrar en esta tarea.

Palabras clave: corpus, investigación basada en internet, *blogs*, vocabulario en lenguas para fines específicos.

Introduction

Curious as it may seem to digital natives – and even to digital immigrants – the use of online sources in academia still causes a certain degree of suspicion and bewilderment among non-specialists. Some recent headlines suggesting this by foregrounding the type of sources used read as follows: "Using Twitter, linguists find global 'superdialects[1]'", "Cambridge University linguists use Twitter to study how Welsh language use is changing[2]", or "Status update language used to predict Facebook users' age, gender, personality[3]".

Precisely because social media and the Internet in general play a crucial role in our world, they tend to record everyday usage as any other medium. This explains why a constantly increasing number of works have concentrated on the potential use of these tools for research (see, amongst others, Zappavigna, 2012; Page et al., 2014; Wang & Winstead, 2016).

Since any form of Internet communication is singular in terms of levels of orality, the amount of visual vs. verbal material employed, and the (a)synchronicity of the medium, the existing diversity enables researchers to select the medium that has the greatest potential to answer the questions we endeavour to explore. In this case, I was interested in examining blog posts, which epitomise asynchronous online communication of a considerable length and with certain – albeit intuitive – textual conventions.

Blogs are usually written by non-experts in the field and infused with a much more personal tone than other digital genres. As Zhang (2008: 37) suggests, "while the Internet has often been thought to decenter the role of the author (Bolter, 1991), blogs have often made the author the center of attention". Still, it is not uncommon to find businesses advertising their services or products via blogs, which, nonetheless, use a narrative style that is considerably less detached from the customer than advertisements simply featuring goods.

As for this particular project, I did not intend to analyse the different narratives of these two subtypes of blogs separately, I focused on the lexical features of both personal and businesses' blogs. My aim was to compile a sufficiently large enough dataset as to be able to trace any patterns, signs of language innovation, and oscillations in the use of native Spanish words *vis-à-vis* foreign words (namely, French-origin and English-origin words), so that I would be able to compare the lexis occurring in blogs to the one used in magazines (online and in print) at later stages. In so doing, my long-term aim is to examine to what extent bloggers are imbued with the discourse present in magazines and repeat the same lexical chunks – namely collocations – and conventions found in these more established genres, or if bloggers diverge from them in order to construct their virtual identities or internet personae. This will hopefully lead me to a deconstruction of this language for specific purposes as realised in these genres. Hence, even though the goal of this study is presently descriptive, once it has been completed, its findings may serve as a theoretical basis for the design of pedagogical material and activities (for example, to assist in the teaching of vocabulary and genre patterns specific to this area of LSP) in the long run.

In the next section I will discuss why blogs are important in our current understanding of languages for specific purposes and how they can contribute to the discipline in comparison to well-established corpora. Subsequently, I will explain the type of data set I compiled, the procedure, and the criteria I followed for their compilation, so that researchers working not only in the area of lexicology but also in other fields of linguistics for specific purposes may draw inspiration and incorporate this 21st-century vast source of data into their practices.

Blogs and 21st-century corpus-driven tools

Whereas the potential of blogs for teaching a foreign language has been extensively explored (*inter alia*, Zhang, 2008; Thomas, 2009; Wang & Winstead, 2016), less has been written on their use in linguistic – more precisely, lexicological – research. However, their functionality cannot be neglected for several reasons. First, blogs contain highly structured information, being typically composed of a homepage (featuring the most recent posts) and an archive. Therefore, the procedure for compiling a blog *a priori* seems to be rather straightforward. Second, although the actual number of posts within each blog varies, the language samples they provide are considerably longer than other virtual genres. It is often argued that a blog should contain at least 300 words to rank well in search engines, although in the database I compiled the total number of words ranged from c. 200 to 700 words. Thus, this virtual space allows writers to devote as much time and space as they feel they need to fulfil their needs with no constraints. Third, the array of non-specialist and specialist voices which can be electronically heard through blog posts create a composite picture which can be used to characterise the speech of these virtual communities.

Fourth, as Hundt et al. (2007: 3) point out, "for the study of certain phenomena, in particular neologisms, the web is and probably will be one of our best sources of information". Like other web-based means of communication, blogs are indeed mines of language data which cover a wide range of topics and are constantly updated. Consequently, lexical items, which are incessantly evolving and being coined, are more likely to make their way into dynamic depositories or databases than into static corpora or dictionaries.

Traditional corpora are finite bodies of text which provide a representative sample of a language either from a diachronic or a synchronic viewpoint. In general Spanish, the most emblematic corpora are *CORDE* (*Corpus Diacrónico del Español*), a diachronic corpus covering the earliest records written in Spanish up until the year 1974, *CREA* (*Corpus de Referencia del Español Actual*), comprising texts from 1974 to 2004, and the most recent one, *CORPES XXI* (*Corpus del Español del Siglo XXI*), which initially incorporated oral and written texts produced from 2001 to 2012 but the project is still underway and will culminate in 2018.

General corpora are expected to include both "standard" and "non-standard" uses of the language so that they present opportunities to study the language in all its forms. However, the asymmetry between the number of "standard" and "non-standard" texts included is problematic when tracing and predicting the development of new "non-standard" uses. For example, in English such a large corpus as *COCA* (*Corpus of Contemporary American English*) only yielded eight tokens of "past tense spreading" with *swim* (i.e. **swimmed*), whereas over 88,000 instances were retrieved from the Web (cf. Geeraert & Newman, 2011: 2). This figure is particularly significant for the language researcher, whose task is to document *any* exiting and emergent uses occurring in the language.

When it comes to languages for specific purposes, particularly, in their didactic facet, *ad hoc* corpora are also developed to examine and teach/learn the features of particular genres and language contexts which are not well represented in general corpora (on this topic, see, for instance, Bárcena, Read, & Arús, 2014). Specialised corpora vary greatly in size and level of specificity, and they may concentrate on very specific text types such as grant proposals (Connor & Upton, 2004) or on the language for specific purposes as a whole (for example, the *Cambridge Business English Corpus*). The dataset that I will broadly characterise in the succeeding section resembles specialised corpora, since the focus is exclusively on fashion blogs written in Spanish, although its relatively limited size (around 900,000 words) compared to well-established corpora prevents me from naming it as such.

English blog-based corpora include the *Birmingham Blog Corpus* (<wse1.webcorp.org.uk/home/blogs.html>), consisting of 628,558,282 words extracted from blogs, and the *Blog Authorship Corpus* (<u.cs.biu.ac.il/~koppel/BlogCorpus.htm>), which encompasses the posts of 19,320 bloggers from <blogger.com> in August 2004, totalling over 140 million words. Nonetheless, these are General English corpora and texts are not categorised into any semantic fields "but split into sections according to how the texts were discovered and downloaded", as the *Birmingham Blog Corpus* website explains. In this sense, the dataset I compiled has a much narrower scope, which may facilitate the task of the researcher interested in describing contemporary Spanish language of fashion and how fashionistas are portrayed in blogs.

Previous research has already concentrated on a limited number of fashion blogs, but mostly from a completely different angle from the enterprise outlined in this paper: for instance, Ruiz Molina (2012) analysed the impact of blogs on consumer-centred companies and fashion journalism (also the main focus of Rocamora, 2012) by adopting a semiotic framework; Riera and Figueras Maz (2012) examined whether fashion blogs attempt to perpetuate the same idealised beauty standard as the one that is promoted in most fashion magazines, something which did not prove to hold true for all of the sub-categories of blogs (e.g. "ego-blogs") equally; along these lines, Rocamora (2011) considered the process of identity construction and representation of femininity in this virtual space; and, more recently, Martínez Navarro and de Garcillán López-Rúa (2016) explored how the emergence and popularisation of this electronic medium has reshaped the practices of consumers through the conduction of interviews and group meetings with informants. On account of this, it becomes clear that the present piece of research is not pioneering on fashion blogs, although its more linguistically oriented aim does offer some scope for originality.

Similarly, even though the use of corpus tools in the compilation and tagging of texts has been widely discussed in LSP research (see, amongst others, Flowerdew, 2005; Millar & Budgell, 2008; Carrió-Pastor & Muñiz-Calderón, 2013; Herrero-Zazo, Segura-Bedmar, & Martínez, 2013) and, to a lesser extent, in thematically diverse e-genres such as blogs (e.g. Wallsten, 2005; Lukač, 2011; Ptaszynski, 2012), substantially less has been written on their interface, that is, on corpus-based approaches to the field of fashion blogs, a gap which the investigation herein presented attempts to fill.

A case study: The fashion blog database

The dataset I assembled consists of the posts of 100 bloggers gathered from 2013 to 2016, amounting to 2,927 blog posts (~900,000 words). Many blogs were located within websites of magazines which count with sections devoted to fashion (for example, <www.telva.com> or <fashion.hola.com/>) although others had independent domains and were found by simple searches containing the key words *blogs*, *moda*, and *España*.

The guiding principles underlying the selection of textual material revolved around the following main axes: (1) language, (2) length, (3) date of production, and (4) topic. A criterion that was essential for the purposes of this project was that all blog posts had to be written in Spanish. It was found that frequently bloggers wrote their posts in both English and Spanish, but this practice would not give us *real* insights into the state of Spanish (for instance, as far as the role of foreign lexis in Spanish is concerned). Accordingly, for the sake of consistency, all blogs had to satisfy this language criterion and be written in Spanish.

Likewise, blogs could not exceed the maximum of 700 words nor could they be shorter than 200 words. In that way, despite variations, I made sure all blogs posts contained a minimum amount of text and were not simply made up of photos. As for the date of production, I concentrated on the time spanned from January 2013 to January 2016. This was a short time span as I intended to pinpoint the most recent lexical trends. However, this could be altered to suit the researcher's needs. The same can be applied to the topic(s) discussed in blogs posts.

I only considered those texts that were within the sphere of fashion, excluding those posts, which, despite being part of a fashion blog, touched upon completely unrelated areas. This thematic decision entirely depends on researchers, their precise aims, and the specialised domains on which they are working (tourism, economics, natural sciences, or medicine, to name but a few).

Once I had retrieved the posts that would be part of my dataset, I created a labelling system in order to identify each file (since each post had an independent file). These labels would include an initialism and a number. These two elements referred to (1) the name of the blog from which they were retrieved (e.g. PAM "Persiguiendo a Mar" <www.telva.com /blogs/persiguiendo-a-mar/> or MDE "Maquillaje de estrellas" <www.mujerhoy.com/belleza/blog-maquillaje-estrellas>), and (2) the date in which the post was released. If more than one post was produced on the same day, I numbered them (e.g. 05-12-2015(1)). For the purposes of this project, I removed all images and only kept written texts so that they could be easily processed electronically.

Furthermore, I recorded basic information of each blog: its (1) name, (2) sub-topic (accessories or clothing, amongst others), and (3) type (personal blog, commercial blog, or others), and (4) the link from which it had been obtained. Depending on the type of study conducted, the researcher can also benefit greatly from knowing the main intended readers, author's background, and other (con)textual elements.

Nowadays electronic tools make it possible to handle large sets of data as these efficiently. There are tools (e.g. *Word Smith* or *Sketch Engine*) which work with any corpus (regardless of its size, type and language(s) in which it is written), and they provide basic features of head-words' behaviour such as KWICs (Key Words in Contexts) and collocations (using Cowie's (1981) and Howarth's (1996) theoretical framework). Therefore, the use of blogs as primary material should not pose any problems in this regard.

It is likewise worth mentioning that, despite the great assets of the online medium, apparently minor issues, such as misspellings, orthographic variations, or foreign words that, for example, have been unconventionally adapted to the Spanish graphemic system, could hinder vocabulary research to a certain extent. This is a pitfall which should be further addressed because not all misspellings can be predicted. Still, if they are recurrent spelling or grammatical errors, they may hint at possible extended patterns that would require further study.

Concluding remarks

This short article has aimed to provide researchers with basic guidelines on how to approach the compilation of online textual material. In particular, I have concentrated on blogs and described the procedure used to compile a real-life linguistic dataset, which may be replicated in almost any other field. Two of the greatest advantages of the Internet is that online users produce language fairly naturally, without being particularly aware that they might be observed for scholarship research – thereby avoiding the observer's paradox –, and that information is constantly renewed and produced unlike paper-based lexicographical sources.

One may argue that blogs will be completely superseded by vlogs or other genres, but whilst this shift may be noticeable in some Internet communities of users, the truth is that blogs still continue to be very much used in such fields as fashion or tourism. Precisely, research into specialised languages requires restricted subsets of language of the type that blogs present. For instance, if one is interested in examining in-group markers used in the language of fashion, it is very likely that fashions magazines will not attest them in the same way as fashionistas or fashion enthusiasts who are writing their own posts with a set of conventions and with a – many times interactive – audience in mind. By compiling a corpus of texts online, not only is it possible to capture myriads of snapshots of the language at that time but also to constantly add new texts which reflect the changing nature of language itself.

All of this will hopefully inspire a much more flexible view on corpus-driven studies of languages for specific purposes. At least ever since 1980 traditional corpora have been used (Heuberger, 2016: 24), so it is time "to keep with up the digital age" and take full advantage of the spectrum of resources that the Internet offers.

Article history:
Received 05 September 2016
Received in revised form 13 December 2016
Accepted 15 December 2016

References

Bárcena, E., T. Read, & J. Arús (2014). *Languages for Specific Purposes in the Digital Era*. New York: Springer.

Bolter, J.D. (1991). *Writing Space: The Computer, Hypertext, and the History of Writing*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Carrió-Pastor, M.L. & R. Muñiz-Calderón (2013). “The compilation of a corpus of business English: Syntactic variation”. *Procedia - Social and Behavioral Sciences* 95: 89-95.

Connor, U. & T.A. Upton (2004). “The genre of grant proposals: A corpus linguistic analysis” in U. Connor & T. Upton (eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics*, 235-256. Amsterdam/Philadelphia: John Benjamins.

Cowie, A.P. (1981). “The treatment of collocations and idioms in learners’ dictionaries” in Peter Strevens (ed.), *In honour of A.S. Hornby*, 223-235. Oxford: Oxford University Press.

Flowerdew, L. (2005). “An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies”. *English for Specific Purposes* 24,3: 321-332.

Geeraert, K. & J. Newman (2011). “*I haven’t drank in weeks*: The use of past tense forms as past participles in English corpora” in J. Newman, H. Baayen & S. Rice (eds.), *Corpus-Based Studies in Language Use, Language Learning, and Language Documentation*, 13-34. Amsterdam/New York: Rodopi.

Herrero-Zazo, M., I. Segura-Bedmar & P. Martínez (2013). “Annotation issues in pharmacological texts”. *Procedia - Social and Behavioral Sciences* 95: 211-219.

Heuberger, R. (2016). “Corpora as game changers: The growing impact of corpus tools for dictionary makers and users”. *English Today* 32: 24-30.

Howarth, P.A. (1996). *Phraseology in English Academic Writing*. Tübingen: Max Niemeyer.

Hundt, M., N. Nesselhauf & C. Biewer (2007). *Corpus Linguistics and the Web*. Amsterdam/New York: Rodopi.

Lukač, M. (2011). “Down to the bone: A corpus-based critical discourse analysis of pro-eating disorder blogs”. *Jezikoslovlje* 12,2: 187-209.

Martínez Navarro, G. & M. de Garcillán López-Rúa (2016). “The influence of blogs in fashion in consumer behavior: An exploratory approach”. *Vivat Academia* 135: 85-109.

Millar, N. & Budgell, B. (2008). “The language of public health - a corpus-based analysis”. *The Journal of Public Health* 16,5: 369-374.

Page, R., D. Barton, J.W. Unger & M. Zappavigna (2014). *Researching Language and Social Media: A Student Guide*. New York: Routledge.

Ptaszynski, M. et al. (2012). “Annotating syntactic information on 5.5 billion word corpus of Japanese blogs” in *Proceedings of The 18th Annual Meeting of The Association for Natural Language Processing (NLP-2012)*, 385-388. New York: Curran Associates.

Riera, S. & M. Figueras Maz (2012). “El modelo de belleza de la mujer en los blogs de moda.¿Una alternativa a la prensa femenina tradicional?”. *Cuestiones de género: de la igualdad y la diferencia* 7: 157-176.

Ruiz Molina, E. (2012). *Blogs de moda: un análisis semiótico*. Barcelona: Design, Knowledge & Future.

Rocamora, A. (2011). “Personal fashion blogs: Screens and mirrors in digital self-portraits”. *Fashion Theory* 15,4: 407-424.

Rocamora, A. (2012). “Hypertextuality and remediation in the fashion media: The case of fashion blogs”. *Journalism Practice* 6,1: 92-106.

Thomas, M. (2009). *Handbook of Research on Web 2.0 and Second Language Learning*. Hershey/New York: IGI Global.

Wallsten, K. (2005). “Political blogs and the bloggers who blog them: Is the political blogosphere and echo chamber”. In *American Political Science Association’s Annual Meeting. Washington, DC September*, 1-36. Online. URL: <journalism.wisc.edu/~dshah/blog-club/Site/Wallsten.pdf>

Wang, C & L. Winstead (2016). *Handbook of Research on Foreign Language Education in the Digital Age*. Hershey/New York: IGI Global.

Zappavigna, M. (2012). *Discourse of Twitter and Social Media*. London: Bloomsbury.

Zhang, F. (2008). *Handbook of Research on Computer-Enhanced Language Acquisition and Learning*. Hershey/New York: IGI Global.

Amanda Roig-Marín holds a BA in English and an MA in Spanish and English as SLs/FLs (both with “Distinction”), and she is currently doing an MPhil in Linguistics at the University of Cambridge. Her main research interests include English lexicology, Spanish-English lexical influences, and historical linguistics. She has published on these areas and delivered papers at national and international conferences.

NOTES

[1] Headline retrieved from <europe.newsweek.com/linguists-find-superdialects-twitter-263622?rm=eu>.

[2] Retrieved from <techcrunch.com/2013/05/29/cymraeg-tweets/>.

[3] Retrieved from <www.wired.co.uk/article/facebook-language-study>.