EPiC Series in Language and Linguistics

Volume 1, 2016, Pages 424–437

CILC2016. 8th International
Conference on Corpus Linguistics

EPiC Language and Linguistics

Japanese L1 speakers blogging in Spanish: motivations, topics and linguistic properties

María del Pilar Valverde Ibáñez
Nara Institute of Science and Technology, Nara, Japan
pilar@is.naist.jp

Abstract

Corpus-based research on learner written language is usually based on corpora made of assignments or exams. In this paper, we instead study personal blogs of learners, since they are a good source of information about learners' motivation to write in a foreign language and their favorite topics, and they have particular linguistic properties. We constructed a blog corpus made up of 2,125 texts coming from 48 Spanish blogs written by 43 Japanese L1 speakers. We find that the main motivations to write a blog are documenting one's life and explaining Japanese culture to a foreign reader. With regard to their linguistic properties, blogs have much in common with both spoken and written language, and the language used is rather informal: we find a higher proportion of lexical verbs, adverbs and personal pronouns and interestingly, it contains also foreign words (Japanese words) and emoticons with various functions.

1 Introduction

Corpus-based research on learners writing is usually based on learner corpora made up of classroom assignments or language level exams. In this paper, we instead focus on personal blogs (short for weblogs[1]) written in Spanish as a foreign language by speakers of Japanese L1, to answer the following questions: What are their main motivations to start a blog in a foreign language? What topics do learners write about? What are the distinctive linguistic properties of this type of texts?

We adopted a corpus-based approach and compiled a corpus made up of 48 weblogs. First, we compiled a list of relevant blogs and used the information on the personal profile to extract information about bloggers' motivations to write. Second, we built a corpus with the texts, and extracted the most frequent topics and the distinctive linguistic properties of texts.

[1] A weblog is defined as a "frequently updated website consisting of dated entries arranged in reverse chronological order so the most recent post appears first" (Herman *et al.*, 2005). In this paper, we refer to the "dated entries" as "posts" or "texts", and to the collection of posts as a "blog" or "weblog". We use the terms "blogger", "writer" and "author" as synonyms to refer to the author of the texts, which in our study is always a leaner.

A.Moreno Ortiz and C.Pérez-Hernández (eds.), CILC2016 (EPiC Series in Language and Linguistics, vol. 1), pp. 424–437

We are interested in blogs because they are written on the learner's own initiative and because of their unique characteristics. First, weblogs, like diaries, give the author the freedom to choose what to write about and in which style. As a result, blogs are good source of information about learners' motivations to write, favorite topics, and even personality (Gil *et al*. 2009).

Second, blogs have much in common with both spoken and written language (Nilsson, 2003). They tend to be an informal genre -the language used in them is less constrained by rules than in academic texts, for example-, but are usually more polished, grammatical and structured than speech or other electronic genres such as tweets or posts on social networking sites.

Third, weblogs are both a monologue and a dialogue (Efimova and de Moor, 2005). Unlike writers of traditional personal diaries which rarely see the light, writers of on-line personal diaries are aware of audience attention, feedback and feelings, so they need to calibrate what they should and should not reveal, and appropriately reply to comments left by readers.

In section 2 we describe the decisions and steps involved in building the corpus, in section 3 we draw some conclusions about the bloggers' motivation to write, in section 4 we summarize the most frequent topics of the texts and in section 5 we examine its linguistic properties of the blog corpus.

2 Corpus Construction

We constructed a Spanish corpus made up of texts written by Japanese L1 speakers. First, we discuss the design decisions taken to select the data sources and then describe the post-processing steps typical of corpora extracted from the Web.

2.1 Blog Search

We searched for written blogs that fulfilled the following requisites:

1. Are written only in Spanish. While some degree of language mixture is acceptable, bilingual blogs which write the same text in two or more languages systematically are excluded.
2. Are written by Japanese L1 speakers. This was confirmed by checking the personal profile or the first posts of the blog. We excluded blogs written by a heritage speaker of Spanish (a person raised in a home where Spanish is or was spoken and who is to some degree bilingual), since their acquisition of the language is different from Japanese native speakers learning Spanish as a foreign language.
3. Are run by a single person. Class blogs are excluded since we cannot determine the identity of the author of the post.
4. Do not have a commercial purpose. Blogs written to attract customers to an on-line business are excluded.
5. Contain genuine text. Blogs containing copies of newspapers news or other sites are excluded.

Such blogs were searched for manually by means of search engines. First, we searched for blogs that contained expressions like "soy japonés", "soy japonesa", "practicar español+Japón", etc. in the bloggers' profile page. Second, by visiting the Blogger's profile page, one can follow any of the links on the page to get a list of other blogs with the same interests (Español, España, スペイン語、スペイン), location, etc. Finally, we visited each blog to check its content and confirm it fulfills the requisites listed previously.

As a result, we obtained a list of 48 blogs written by 43 learners, hosted mainly in Blogger and Wordpress domains. By visiting the bloggers' personal profile, we were able to assign gender, location and other personal information to each of them. Two thirds of bloggers (29) are women and one third (14) are men. Almost half of the bloggers are living in Spanish-speaking countries (16 in Spain, 3 in Mexico and 1 in Costa Rica), to study Spanish as undergraduate students (for a short period of time) or they are residents of the country for other reasons, usually for a longer period of time (from 2 to 18 years). Of those living in Japan (23), several have visited or have lived before in a Spanish-speaking country, too. The oldest blog started in 2004 and the most recent one in 2015. The most active blogs started between 2010 and 2012.

2.2 Data Collection and Post-Processing

From the 48 blogs we extracted a list of URLs, one for every post, totaling 2,701 texts. Using the `wget` utility[2], in the beginning of 2015 we downloaded the corresponding HTML files that were post-processed to obtain the plain text corpus to obtain, for every blog post, a plain text file containing the text written by the blogger

By post-processing, we refer to "non-linguistic cleanups which are required to turn the collection of downloaded HTML documents into a collection of documents ready to be included in a corpus" (Schäfer and Bildhauer, 2013). This involved cleanups within the documents and the removal of documents which did not meet certain criteria. HTML files were post-processed in the following steps:

1. Conversion of files to UTF-8 encoding. Each downloaded document was checked for its encoding and, if necessary, converted to UTF-8 with the `iconv` tool[3].
2. Selection of body text. We were interested only in the part of the file that contains the text written by the learner (body text), so we discarded the header and the comments sections. To do that, we used regular expressions that take advantage of HTML structure markup. Although blogs may look different in their design and contain different content, the underlying structure is very similar. For example, the main body of text in Wordpress posts often appears after the tag (with any number after "post-") and before the tag .
3. HTML stripping. We used the `textutil` tool[4] to exclude HTML tags, figures and other multimedia content, and leave only plain text.
4. Boilerplate removal. Texts still contained some irrelevant material that was removed by means of regular expressions:
    - Redundant content like navigational elements (*Archives, Categories, Email, Recent Posts, Search for*), linked content from the same category (*Recent entries, Post Relacionados*), suggestions (*Subscribe, Deja un comentario, Leave a Reply*) or blog titles.
    - URLs and email addresses, which are not linguistically relevant and should be removed in order to protect privacy.
5. Text size selection. Texts with less than 30 words were excluded.

[2] https://www.gnu.org/software/wget/
[3] https://www.gnu.org/software/libiconv/
[4] https://developer.apple.com/legacy/library/documentation/Darwin/Reference/ManPages/man1/textutil.1.html

6. Language identification. Even though we only selected monolingual blogs, some bloggers occasionally write some posts in another language or the same post in two or three languages. We used the `ngramj` tool[5] to exclude files in other languages and files with less than 60% of Spanish content.

The number of documents resulting from this processing was 2,125, as shown in Table 1.

Post-processing step	Texts
Downloaded HTML files	2,701
UTF-8 encoding conversion	2,700
Selection of body text	2,685
HTML stripping	2,682
Boilerplate removal	2,177
Text size selection	2,177
Language selection	2,125

Table 1: Number of texts after each post-processing step

The corpus finally contains 737,396 tokens, 625,343 words, 45,787 sentences and 2,125 documents. It was uploaded to the Sketch Engine corpus query tool[6](Kilgarriff *et al.* 2014) and automatically part-of-speech tagged with Spanish Freeling v4 (Padró and Stanislovsky, 2012) to explore its linguistic properties in more detail.

3 Bloggers' Motivations to Write

Motivations for writing a blog are varied and several classifications have been proposed in the literature (Li 2005 and Nardi *et al.*, 2004). Interestingly, the most frequent motivation of our bloggers is not to improve their language as one would expect given their condition of non-native speakers. Only three blogs explicitly state this is their main motivation, although many more may indirectly have this goal. Taking into account that different motivations might happen simultaneously, in our Spanish blogs we find, according to what they say in their personal profiles, that bloggers write, (in order of descending frequency) to document their life (section 3.1), explain Japanese topics to foreign readers (3.2), improve their writing (3.3), discuss external topics (3.4) and socialize (3.5). Then, their two main motivations are first to document their lives (20/43), as in a personal diary, and second, to explain topics related to Japanese culture to readers from Spanish-speaking countries (17/43).

Consistent with the idea that woman tend to be personal blog authors while men are more motivated to transmit information (Li 2005), in our data the motivation of documenting one's life is slightly more frequent among women (48,3%) than among men (35,7%), while providing information (about Japanese culture or external topics) is slightly more frequent among men (50,0%) than women (34,5%).

3.1 Document One's Life

Often blogs function as a combination of diary, notebook and mailing list. Bloggers concentrate on mundane topics concerning personal experiences and emotions and personalize the look of blogs to

[5] http://ngramj.sourceforge.net/

[6] https://www.sketchengine.co.uk/ Users of this tool can have read-only access to the blog corpus for research purposes. Please contact the author of this paper if you would like to have access to the corpus.

make it part of their self-presentation. Blogs can be used as a notebook, to keep record of what the blogger learns or thinks (1). Students living abroad may want to keep a memory of their experience abroad, and those who already returned to Japan may use the blog to keep in touch with their Spanish-speaking friends (2). Bloggers living in a foreign country for a long time may also use writing as a catharsis (3).

(1) Gracias por visitar aqui!soy japonesa y vivo en Osaka.Escribo mi vida para no olvidarla y practicar espanol.スペイン語なんとかせんとー！と日々足掻いています。

(2) Gracias por leer mi blog que apenas escribo. Empecé este blog para practicar español (agradezco cualquier comentario para mejorar mi español). También me gustaría que mis amigos supieran algo de mi, que sigo andando por aquí, recordando vosotros de vez en cuando.

(3) Hola!! Soy Nao!! Durante 5 años yo estaba en Centroamérica y me casé con un tico y tengo una niña preciosa que nació en el 2008! :D Escribo lo que siento a través de mi vida.

3.2 Explain Japanese Topics to Foreign Readers

Bloggers gain satisfaction by providing information to others for whom it may be difficult to obtain such information. Japanese learners feel they have privileged knowledge about Japan which their readers do not have, so they behave as intermediaries between Japanese culture and Hispanic culture, trying to explain Japanese culture to a Spanish-speaking reader (4, 5, 6).

(4) A partir de hoy, voy a empezar este blog para que muchos hispanohablantes conozcan y tengan interés a mi país.

(5) soy japonesa, vivo en Madrid. al ver la situacion de que los españoles estan buscando la manera eficaz y adecuada para controlar su cuerpo y mantener la salud corporal al igual que mental, decidi a presentar el truco de bienestar y belleza de las japonesas, que todavia no se conoce muy bien en España.

(6) Hola, en mi blog, te voy a presentar la música japonesa que apreciamos los japoneses, que nos acompaña en nuestra vida cotidiana y que ha tocado nuestro corazón!

3.3 Improve their Writing

Some bloggers hope that they can improve their language skills by writing and interacting with their readers (1, 2, 7, 8, 9). Since their texts are public, they sometimes expect others will point out their mistakes.

(7) Soy tokiota. Me gusta escribir frases en español. Me alegraría que te encantara lo de mi blog.

(8) Soy OTAKU:) de Japon. Tengo 23 años. Soy estudiante de Japón y ahora estoy estudiando español en España hace un raro. Este blog hay para practicar en escribir en español. Y voy a escribir sobre Japón, Amime, Game y Vovaloid(Niko Niko Douga). Mi español es muy muy muy muy muy muy malo. Intento escribir para que todo el mundo lo entienda! Gracias por leer. Encantada:)

(9) Soy japonés y estoy pensando emigrar a Sudamérica.Hago este blog para que me aprenda e. Por eso hay muchas faltas de las gramáticas, las expresiones, las ortografías, etc.español.

3.4 Discuss External Topics

Some blogs discuss external topics of interest to the author, and express their opinions and feelings about them. Bloggers can be unofficial eye-witness to news stories, such as the two blogs in this category, which started on the occasion of 2011 Fukushima nuclear disaster (10, 11).

(10) Soy Natural de Iwaki (Fukushima, la provincia más afectada por el terremoto) y voy a publicar mis pensamientos, noticias y opiniones de personas que han sufrido y sufren esta pesadilla en mi pueblo

(11) Soy Japonesa. Vivo en España. ¿Quieres energía nuclear? No, gracias.

3.5 Socialize

Since blogs make possible to interact with readers or bloggers who share similar interests, they can be a platform to develop interpersonal relationships (12).

(12) Tengo 17 años, soy Japonesa pero vivo en España. Quiero hacer amigos.

4 Blog topics

Web blogs give the writer freedom of topic and style but, unlike in traditional personal diaries which remain confidential, writers need to calibrate what they should and should not reveal, considering readers' feedback, background and feelings. In these conditions, what do Japanese L1 speakers write about in Spanish?

The blogs in our study deal with the following topics: personal life (23), Japanese culture (12), food and recipes (8), languages (3) and Fukushima nuclear disaster (2). This is consistent with the bloggers' motivations seen in section 3. Bloggers write mainly about their daily life (in Japan or in a Spanish-speaking country), Japanese culture, Japanese or Spanish language and a current national problem like Fukushima nuclear disaster.

The bloggers' interests are reflected in the most frequent nouns in the corpus, shown in Table 2. It is interesting the high frequency of temporal nouns like *año* (*este año, el año pasado, llevo años viviendo en, hace años*, etc.) *día* (*ese día, cada día, todos los días, muchos días*, etc)., *vez* (*la próxima vez, muchas veces, por primera vez, a veces*, etc.) and *tiempo*, which can be attributed to the narrative style of the texts dealing with daily life.

Bloggers also frequently write about Japanese culture, and compare it with Hispanic culture, as is shown by the high frequency of words like *Japón, casa, vida, gente, España, país, español, mundo, cosa*. Japanese food, in the form of recipes (*persona, arroz, comida, agua*) is especially well represented, as show the most frequent object collocates of (*no*) *gustar*: *sabor, comida, cocinar, comer* are more frequent than *leer ver, hablar, viajar, ir*; and *japonés, español* are its most frequent subjects. The most frequent collocates of *japonés* are *comida, tienda, cocina, cultura, estilo, tortilla, televisión, amigo, té, restaurante, costumbre, chico, salsa, postre, idioma*. And the most frequent collocates of *español* are verbs like *hablar, estudiar, aprender, dominar, entender* and nouns like *clase, nivel, academia, mayoría* (*de*).

Lemma	Count
japón	1,894
fukushima	1,887
año	1,779
vez	1,660

día	1,529
casa	1,199
persona	1,172
vida	1,002
arroz	944
pueblo	940
tiempo	936
gente	912
españa	911
país	855
español	807
comida	805
mundo	803
cosa	751
kimono	700
agua	670

Table 2: Top 20 most frequent nouns (lower-case lemmas) in the corpus

5 Linguistic Properties of Texts

Herring *et al.* (2005) and Nilsson (2003) have showed that the language of blogs has much in common with both spoken and written language. Blogs do not belong completely to either medium, written or spoken, but instead fall somewhere in between.

Like in writing, blogs 1) are space bound -text is bound to the space it occupies- and asynchronous, 2) can employ constructions which are characteristic of writing -like specialized vocabulary and complex sentences- and 3) can be revised, both before and after they are published (although it is not common to do so).

In addition, like in speaking, blogs 1) are dynamic -the front page, containing the most recent posts, changes often, while the older posts are kept archived-, 2) promote immediacy -texts are published shortly after the event they narrate occurs-, 3) attempt to express speech nuances -by means of emoticons- and 4) employ constructions which are characteristic of speech (slang, contractions, nonsense words, etc.).

In our corpus, we find a higher proportion of word classes characteristic of informal texts -personal pronouns, lexical verbs and adverbs- (section 5.1), foreign words (5.2), emoticons (5.3), constructions characteristic of speech (5.4) and constructions characteristic of writing (5.5).

5.1 Word Classes Characteristic of Informal Texts

In several languages, nouns, adjectives, articles and prepositions are more frequent in formal texts, while personal pronouns, adverbs and lexical verbs are more frequent in informal styles (Heylighen and Dewaele, 1999). Indeed, our corpus contains a higher proportion of personal pronouns, lexical verbs and adverbs than other types of texts -written by learners or by native speakers-.

As shown in Table 3, lexical verbs represent 14.8% or words, adverbs 5% and personal pronouns 2.7%. These figures are slightly but consistently higher than in other three corpora: a corpus of academic texts written by Japanese university students of Spanish as a foreign language (APU)

(Valverde, 2015), a corpus of academic texts written by native university students in México (UNAM) (sub-corpus of CLAE, 2009), and the esTenTen Spanish Web corpus (Jakubíček *et al*., 2013).[7]

Type of writer:	Learner				Native			
Type of text:	Blog		Academic		Academic		Web	
Corpus name:	Blog corpus		APU		UNAM		esTenTen	
Frequency	a.f.	%	a.f.	%	a.f.	%	a.f.	%
Lexical verbs	92,306	**14.8**	7,307	11.6	29,648	11.5	1,155,935,910	12.2
Adverbs	31,341	**5.0**	1,968	3.1	7,917	3.1	285,060,011	3.0
Pers. Pronouns	16,648	**2.7**	630	1.0	2,750	1.1	120,290,402	1.3
Total words in the corpus	625,343	100	63,069	100	258,597	100	9,497,402,122	100

Table 3: Frequency of lexical verbs, adverbs and personal pronouns in learner and native corpora.

5.2 Foreign Words

Code-switching, that is, alternating two or more languages in bilinguals' discourse, is common in blogs written by bilingual speakers (Montes-Alcalá, 2007). Our blogs do not contain proper code-switching -we explicitly excluded blogs written in two languages or from heritage Spanish speakers from our analysis- but we frequently find words and expressions in the writers' first language, Japanese, sometimes next to its Spanish equivalent, with different functions.

Japanese words are mainly used for expressing culture bound concepts. Therefore, the occasional use of Japanese words inside the Spanish texts should not be interpreted as a lack of language proficiency, instead, as a result of the lack of an exact equivalent in the target language. Our bloggers are in between two cultures and two languages, and they need to use both of them to fully express themselves.

As a downside, the combination of Japanese and Spanish words involves the use of four different scripts, which complicates the automatic treatment of texts: the Latin script, used to write Spanish (e.g. *alga*) or to transcribe Japanese (in lower-case e.g. *konbu* or upper-case e.g. *KONBU*), Japanese kanji (logographic Chinese characters, such as 昆布), hiragana (だし) and katakana (ウルトラマン).

Although Spanish words are mainly written using the Latin alphabet, we frequently find Japanese characters, specially punctuation marks[8] and even –probably unintentional- awkward combinations of full-width and half-width characters in the same sentence.[9]

It is interesting that bloggers include the word in kanji or kana in their texts (usually next to its transcription in the Latin alphabet), given the fact that readers probably cannot read kanji and kana. The inclusion of all the written forms of a word can probably be explained by the dual nature of blogs: since they are at the same time personal and public diaries, their text must be pleasant to read for both the blogger (who prefers to use the Japanese writing system for Japanese words) and the reader (who needs its transliteration).

[7] All four corpora have been assigned part-of-speech tags automatically with Freeling v4 (Padró & Stanislovsky 2012). Although automatic annotation inevitably implies a certain error margin, which is slightly higher in learner texts (Valverde 2011), we believe that using the same tagging method for the four corpora makes results enough comparable.

[8] Quotes (“”), dots (。), three dots (…) , parenthesis, special characters like ☆, etc.

[9] For example: “Ｅｌ　ｔｅａｔｒｏ　en la parte de atrás de mi es el sitio que bailar　el Nou.”

The use of Japanese can have different functions (in decreasing frequency): refer to Japanese things or concepts (5.2.1) or language (5.2.2), quote somebody's words (5.2.3), add emphasis (5.2.4) and code-switching (5.2.5). In the following examples, Japanese words -written with Japanese kanji or kana or transcribed in the Latin alphabet- are shown enclosed and the equivalent Spanish word or paraphrase is underlined.

5.2.1 Japanese Things or Concepts

The most common is to use Japanese words to refer Japanese things or concepts for which there is no good translation, for example Japanese food or ingredients in a recipe (13), inventions or products (14), festivities or traditions (15) and proper names -people (16), entities (17) or places (18)-.

(13) Lo más importante para la comida japonesa es "Dashi だし".

- ¿Qúe es Dashi?
Es un caldo japonés. Usamos para casi todos los platos japoneses.
=== Ingredientes ===
- Alga (se llama KONBU 昆布) 20 gr. (unos 10 cm. más o menos)
- Bonito seco (se llama KATSUOBUSHI 鰹節) 30 gr.
- Agua 100 ml. y 1000 ml.

(14) Por último, en Japón existe un sentimiento de vergüenza a que los demás escuchen los sonidos que se producen mientras está utilizando el water. Así existe un sistema llamado OTOJIME(REINA DEL SONIDO). Si pulsas este botón, mientras utilizas el water, se produce un sonido de corriente de agua durante unos 15 segundos.

(15) Dentro de un mes, el 5 de mayo, hay una fiesta para los niños en Japón.Se llama "Día de los niños (kodomo no hi こどもの日)".

(16) Pero en esta semana he encontrado una poesía. Se llama 「朝のリレー」(Relevos de la mañana), por 谷川俊太郎 (Tanigawa Shuntaro) Cuando leía una revista japonesa que me mandó una amiga mía, encontré un artículo del poeta y poesía.

(17) Quería hacer mi donación para Fukushima o para los niños que habían perdido a su padre. Entonces decidí donar a あしなが育英会(La Fundación Ashinaga). Ya he donado un poco desde mi cartera.

(18) Conoces Fukuoka(福岡)? Conoces "Los peñascos del matorimonio(夫婦岩 :Meotoiwa)"?

5.2.2 Japanese Language

Japanese words are also used in metalinguistic contexts, when talking about the Japanese language itself, in the form of words (19) or proverbs (20).

(19) ナウい(Naui) Adjetivo que significa "a la moderna". Procede de la palabra inglesa "now" y "i" que es el final de una palabra para hacer un adjetivo japonés. ej) "Su vestido es muy Naui!!" El japonés que usa esto es la persona de la mayor edad.

(20) En japonés hay una frase hecha "Ocha wo Nigosu (literalmente, "enturbiar el té")", que significa "engañar".

5.2.3 Quotes

Japanese words are also used to quote somebody's words, in direct (21) or indirect (22) speech.

(21) Cuando terminamos, me levanté y dije "muchas gracias por la entrevista 本日はお時間をお取りいただきありがとうございました", haciendo una reverencia. Y otra vez delante de la puerta hice otra reverencia.

(22) Según la publicación por la Cruz Roja de Japón, al 16 de mayo, se juntaron las donaciones de 1,925 億 6,316 万 9,033 円 (192.563.169.033 yenes) que equivalen a 1.673 millones de euros para asistir a los damnificados por el gran terremoto y el tsunami del 11 de marzo.

5.2.4 Emphasis

Bloggers can give emphasis by repeating the same sentence in both languages (23), often at the beginning of the post -in the title (24)-, or at the end -in the closure (25)-.

(23) Es la hora de decir muchas gracias, ARIGATOU GOZAIMASU a miles de personas que nos han ayudado y nos están ayudando continuamente.

(24) mi cumpleaños 私の誕生日

(25) Qué durmais como marmotas. ぐっすりおやすみください

5.2.5 Code-switching

Certain linguistic routines or idiomatic expressions are especially frequent at the end of sentences (26, 27, 28). The writer does not provide the translation of these words, so this can be considered as an example of code-switching, where biculturalism seems to plays an important role.

(26) Paulaaa, gomennn tambien estan tus fotos de la fiesta de tu cumpleanos :D omatase…

(27) ¡Es muy curioso y me encanta! KAWAII~~~~~~~~!!!!!

(28) Una amiga de Blugaria hizo fotografias.Bien, ne^^

5.3 Emoticons

Emoticons are used frequently in the corpus to express the nuances of speech. As expected, we find vertical emoticons (29) as well as a large variety of Japanese style horizontal emoticons (30, 31, 32), which make use of a larger character selection, and even other ASCII objects (33).

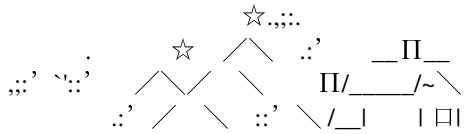
(29) Mi hermano mide 198cm, y lo come desde por la mañana XD

(30) Tanto tiempo no he escrito!! Perdon por vaga!!! m(_ _)m

(31) Hoy era 2años de nuestro matorimonio.(^^)Y☆Y(^^)

(32) +。:.゜ヨロシクーッ!!ヽ(´∀`)ノ.:。+゜。

(33)



5.4 Constructions Characteristic of Speech

The informality of blogs is revealed by the use of slang such as *tío* (34), *chaval* (35), *guay*, *chulo* (36), *molar* (37), *cojones* (38), *cabrón* (39), etc. Interjections (40, 41) and unfinished sentences (42) are also present.

(34) Un tío que había viajado fuera de Japón y le habían gustado los baños arabes montó una sauna en Ginza donde te podías meter en una sauna privada y unas chicas te daban masaje.

(35) Es un chaval guapito con pinta de punk, y la cara llena de agujeros y muchos pinchos en la ropa.

(36) Hoy en día , está de la moda el kimono entre los jovenes de Japón, hay muchos accesarios del kimono, y la manera de vestir el kimono también es más chula y guay como con calcetines (TABI), las botas , la estola y la mochila ...

(37) Me acuerdo que un anyo, antes de florecer empezo sacar miles de orugas. Dia tras dia se iban aumentando sus hijos o primos o sus colegas que algunos se colaban al borde de la casa. En principio las captulaba pero no me molaba tanto como coger otros especies.

(38) Fue la primera vez que organice la fiesta de la boda para mi,pero fue una fiesta de cojones gracias a mis amigotes.

(39) Pero el muy cabrón apareció a casa con una botella.

(40) Hoy, hace buen tiempo. Esta bien. Esta mañana he hecho la colada cuarto veces. ¡Uf! Estoy un poco cansado.

(41) Al final, los novios no tiran nada a amigos en la boda como el ramo ni la liga. jajaja.

(42) Ya no está abierto la tienda ni super.....Las comidas que tenia en la casa son sólo cerdo, pasa, nueces, naranja..... Pues por eso he añadido todo con un poco de aceite de oliva y la mantequilla.

5.5 Constructions Characteristic of Writing

Despite the informal style of most texts, blogs also allows the use of constructions characteristic of writing -such as the relative *cuyo* (43)- and specialized or formal vocabulary when the topic requires it (44, 45). Unlike in speech or more immediate forms of communication, bloggers can revise their text if they need to do so, and use reference materials such as dictionaries and grammars to write their text. We can also find awkward combinations of words typical of speech and writing in the same sentence (46).

(43) Hoy os hablaré de “Kagami biraki” (鏡開き) es una ceremonia tradicional japonesa cuyo nombre significa literalmente “abrir el espejo”.

(44) Se discutirán los poderosos del tema de todo el planeta sobre la descontaminación y el seguimiento de la salud de la población.

(45) Con esa actitud infantil los radioelementos contagiarán al agua del rio, ensuciarán el ecosistema y finalmente llegarán al mar del globo.

(46) Como no conocía nada de español cuando compré mi móvil, elegí uno con el que fuera posible hacer y recibir llamadas y enviar mensajes de texto SMS. Por consiguiente , el diseño es súper simple.

6 Conclusions

Blogs are an interesting are of research because of their ambivalent nature: they are private and public, speech and writing, monologue and dialogue at the same time. In this paper, we have used a learner Spanish blog corpus to examine Japanese bloggers' motivations to write, favorite topics and the linguistic properties of their texts. We have made a careful selection of the blogs that make up the corpus, 48 blogs written by 43 authors. From manual inspection of the bloggers' profiles, we found a majority of women bloggers, and roughly the same proportion of bloggers living in Japan or in Spanish-speaking countries. The two main motivations to write a blog are documenting one's life, as in a conventional personal diary, and two explain Japanese culture (food, customs, traditions, etc.) to a foreign reader. However, the proportion of texts dealing with Japanese culture is smaller than the proportion of texts dealing with Fukushima nuclear accident, since one of the two blogs about this topic is the most productive one in the corpus. Therefore, more than half of the texts deal with daily life, one fourth with the nuclear accident and the rest about Japanese culture and food.

To build the corpus from the list of blogs, we downloaded all the entries (2,701) in every blog and carried out some post processing to select only the main body of each entry and discard irrelevant material. Finally, our corpus, containing 2,125 texts and 625,343 words, was uploaded to the Sketch Engine corpus query tool and automatically part-of-speech tagged.

With regard to the linguistic properties of the texts in the blog corpus, we have confirmed that blogs possess spoken and written traits. As in speech, blogs contain a higher proportion of lexical verbs, adverbs and personal pronouns than other type of texts written by learners or by native speakers. A particular characteristic of our corpus, written by bloggers who have contact with two very distinct languages and cultures, is the frequent use of foreign words (Japanese words) inside the Spanish text, with different functions (referring to Japanese things or concepts, explaining Japanese language, quoting, emphasizing and code-switching). Emoticons are also frequent to express emotions. Finally, we can find linguistic elements characteristic of speech (slang, interjections, unfinished sentences, etc.) as well as elements characteristic of writing such as specialized vocabulary.

As future lines of research, we would like to evaluate our corpus and make it more balanced with regard to the number of topics and bloggers' productivity. It would also be interesting to investigate the differences among learners with different L1, as well as the further linguistic processing of the text so that we can extract data not only about learner vocabulary but also about syntactic and discourse structure.

Acknowledgments

This research was supported by *kakenhi* (25770207), Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science.

References

CLAE (2009), *Corpus de lenguaje académico en español*. UCMexus-CONACYT. Available at http:// www.lenguajeacademico.info

Efimova, L., and de Moor, A. (2005). Beyond personal webpublishing: An exploratory study of conversational blogging practises. *Proceedings of the 37th Annual HICSS Conference*. Big Island, Hawaii.

Gill, A.J., Nowson, S., and Oberlander, J. (2009). What are They Blogging About? Personality, *Topic and Motivation in Blogs*. Association for the Advancement of Artificial Intelligence

Herman, D., Jahn, M., and Ryan, M.-L. (eds.). (2005). *The Routledge Encyclopedia of Narrative Theory*. London, Routledge.

Herring, S., Scheidt, L., Sabrina Bonus, S., and Wright, E. (2005). Weblogs as a bridging genre. *Information, Technology & People*, 18 (2), 142-171.

Heylighen, F., Dewaele, J.M. (1999) *Formality of language: definition, measurement and behavioral determinants*. Technical report, Free University of Brussels.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., and Suchomel, V. (2013) The TenTen Corpus family. *7th International Corpus Linguistics Conference*, Lancaster.

Kilgarriff, A., Baisa V., Bušta, J., Jakubíček,M., Kovář, V., Michelfeit J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on. In Lexicography: 1–30.

Li, D. (2005). *Why do you blog: A uses-and-gratifications inquiry into bloggers' motivations*. Masters Thesis, Marquette University.

Montes-Alcalá, C. (2007). Bloggin in two languages: code-switching in bilingual blogs. *Selected Proceedings of the Third Workshop on Spanish Sociolinguistics*, 162-170. Somerville, MA: Cascadilla Proceedings Project.

Nardi, B.A., Schiano, D.J., Gumbrecht, M., and Swartz, L. (2004). Why We Blog. *Communications of the Association for Computing Machinery*. December. Pp. 41-46.

Nilsson, S. (2003). *The function of language to facilitate and maintain social networks in research weblogs*. Dissertation, Umea Universitet, Engelska lingvistik.

Nowson, S. (2006). The Language of weblogs: a study of genre and individual differences. Doctoral Thesis, The University of Edinburgh

Padró, L. and Stanilovsky, E. (2012) FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)* ELRA. Istanbul, Turkey. May, 2012.

Schäfer, R. and Bildhauer, F. (2013), *Web Corpus Construction*, Synthesis Lectures on Human Language Technologies, Morgan&Claypool Publishers.

Valverde, M.P. (2011). An evaluation of part of speech tagging on written second language Spanish. In Gelbukh, A. (Ed.), *Lecture Notes in Computer Science*, vol. 6609, Springer Berlin Heidelberg, pp. 214-226, ISBN 978-3-642-19399-6 & 978-3-642-19400-9, DOI 10.1007/978-3-642-19400-9_17.

Valverde, M.P. (2015), Frecuencia de uso de palabras gramaticales en textos académicos: Comparación de un corpus aprendices de ELE con tres corpus de referencia. In *Hispánica*, pp. 127-154, Japanese Association of Hispanists, ISSN 0910-7789 & 1884-0574.