ISSN 0214-0489

ELABORACIÓN DEL CORPUS DEL PROYECTO «DICO CULTUREL» (DICCIONARIO CULTURAL): FUENTES FIABLES Y NO FIABLES

ASSEMBLING A CORPUS OF INFORMATION FOR THE "DICO CULTUREL" (CULTURAL DICTIONARY) PROJECT: RELIABLE AND UNRELIABLE SOURCES

Olga Rocío Serrano Carranza
Universidad ECCI. Bogotá, Colombia

Proceso editorial
Recibido: 24/11/2017
Aceptado: 28/11/2017
Publicado: 1/12/2017
Contacto
Olga Rocío Serrano Carranza
oserranoc@ecci.edu.co

CÓMO CITAR ESTE TRABAJO | HOW TO CITE THIS PAPER

Serrano Carranza, O. R. (2017). Elaboración del corpus del proyecto «Dico Culturel» (Diccionario Cultural): Fuentes fiables y no fiables. *Revista de Educación de la Universidad de Granada, 24:* 227-250.

ELABORACIÓN DEL CORPUS DEL PROYECTO «DICO CULTUREL» (DICCIONARIO CULTURAL): FUENTES FIABLES Y NO FIABLES

Resumen

En el marco del proyecto de investigación «Dico Culturel» buscamos describir y analizar palabras y expresiones de la lengua francesa ricas en contenidos culturales antropológicos, estableciendo paralelos lingüísticos e interculturales con respecto al español de Colombia. Para tal propósito, vamos a construir un corpus comparable en estas lenguas fundamentado en documentos auténticos disponibles en Internet. Pero en este punto nos surge un interrogante. El de saber si dicho corpus se conformará sólo de informaciones encontradas en fuentes fiables o también en fuentes no fiables. Para despejar esta duda, estudiamos una muestra de 19 palabras y expresiones de la lengua francesa y a partir de estas constituimos dos corpus en francés: un corpus de informaciones sacadas de fuentes fiables y otro, compuesto de informaciones de fuentes no fiables. Luego, y gracias a un programa de análisis textual, consideramos y comparamos las frecuencias, las concordancias y los contenidos lingüísticos y culturales de las palabras y expresiones de la muestra en cada uno de los corpus con el fin de examinar su calidad y pertinencia. Tras análisis de los contenidos, se concluye que muchos de ellos no presentan diferencias en el corpus de fuentes fiables y en el corpus de fuentes no fiables. Adicionalmente, se observó que múltiples datos se complementan ya que se encuentran o no en uno de los dos corpus, lo que nos lleva a constatar que las informaciones halladas en las fuentes no fiables pueden ser de interés para nuestra investigación sin dejar de lado su verificación con las fuentes fiables.

Palabras clave: Corpus; fuente de información fiable; fuente de información no fiable; contenidos lingüísticos y culturales.

ASSEMBLING A CORPUS OF INFORMATION FOR THE "DICO CULTUREL" (CULTURAL DICTIONARY) PROJECT: RELIABLE AND UNRELIABLE SOURCES

Abstract

In the framework of the "Dico Culturel" research project, we aim to describe and analyze French words and expressions which are rich in cultural and anthropological contents and establish linguistic and intercultural parallels with the Spanish used in Colombia. Towards that end, we are going to construct a corpus which allows for cross-comparisons between these two languages, drawn from authentic documents available on the internet. But this approach raises a question. It is how to know whether that corpus shall only consist of information found in reliable sources or also include the use of unreliable ones. To clear up this doubt, we studied a sample of 19 words and expressions in the French language and on that basis, created two corpora in French: one of information drawn from reliable sources and the other from unreliable sources. Then, with the use of a program of textual analysis, we looked at and compared the frequencies, concordances and linguistic and cultural contents of the words and expressions in the sample of each corpus in order to de-

termine their quality and relevance. After analyzing the contents, we concluded that in many cases there were no differences between the corpus of reliable sources and the corpus of unreliable sources. In addition, we noticed that many data complemented each other, since they were found in one or the other of the two corpora, which led us to confirm that the information found in the unreliable sources may be of interest for our research, without excluding their verification in the reliable sources.

Keywords: Corpus; reliable source of information; unreliable source of information; linguistic and cultural contents.

INTRODUCCIÓN

Este artículo se inscribe dentro del proyecto de investigación «Dico Culturel» cuya finalidad es desarrollar un modelo de diccionario comparativo entre el francés de Francia y el español de Colombia. Se desean englobar las divergencias lingüísticas, culturales e interculturales antropológicas, totales o parciales, entre estas dos lenguas y culturas.

Los contenidos que sustentan el corpus, para la redacción de los artículos de este diccionario, se obtienen primordialmente vía Internet. Al tratarse de datos ya digitalizados se escudriñan más rápida y eficazmente con la ayuda de herramientas informáticas, lexicométricas o textuales.

Ahora bien, al explorar el material bibliográfico accesible en la Web, es imperioso corroborar la calidad de las fuentes de información, su exactitud y veracidad. No obstante, resulta complejo juzgar los documentos de la Red en fuentes fiables o no fiables debido a su gran heterogeneidad, a la falta de normas que viabilicen su categorización y a la facilidad o a la celeridad en la que son publicados.

Nos planteamos por ende las siguientes preguntas: ¿qué tipo de textos e informaciones digitales podrían ser tenidos en cuenta para la constitución del corpus de nuestro proyecto de diccionario cultural? ¿El corpus estaría solamente integrado de datos sacados de fuentes fiables? ¿Podrían ser las fuentes no fiables de utilidad en la búsqueda de informaciones lingüísticas y culturales?

Para responder a estos interrogantes, llevamos a cabo un estudio que es el que exponemos en este artículo. Este consistió en crear dos corpus en francés basados en una misma muestra de palabras y expresiones francesas. El primero de ellos se compone de fuentes fiables y el segundo de fuentes no fiables.

Posteriormente, con un programa de análisis textual, indagamos sobre las palabras y expresiones examinadas, estableciendo paralelos entre los dos corpus en lo

tocante al número de frecuencias, a las concordancias, a los sintagmas complejos conexos, a los contenidos lingüísticos o culturales encontrados y a la calidad de los mismos.

A partir de este análisis y de los resultados conseguidos, llegamos a una serie de conclusiones que comentamos al final. Pero antes, enunciamos los conceptos claves que a nuestro parecer son relevantes para comprender mejor el problema de investigación.

MARCO TEÓRICO

El primer concepto que mencionamos es el de corpus. Un corpus es una suma de textos recopilados según criterios previamente delimitados en formatos, registros y géneros diversos (Flowerdew, 2012, p. 1; Deronne, 2011, p. 42). A través de este se escrutan fenómenos lingüísticos (Müller y Gjerstad, 2014, p. 49; Sinclair, 1994, p. 2, citado por McEnery, Xiao y Tono, 2006, p. 4), culturales, sociales o pragmáticos (Flowerdew, 2012, p. 1). El corpus debe ser representativo (Terrádez, 2001, p. 19) y adecuarse al objeto de la investigación (Olivier, Moré y Climent, 2008, p. 104).

Para formar un corpus se recurre a textos auténticos impresos (Côté y Troudi, s.f., p. 1) o procedentes de Internet (Austermühl, 2014, p. 124; Drago, 2009, p. 240). Estos últimos son vistos por muchos como recursos provechosos e inagotables de datos (Mehler, Sharoff y Santini, 2010, p. 16) dispuestos de manera escrita, visual o sonora (Clarenc, 2011, p. 201; Isaac, Hamon, Fouqueré, Bouchard y Emirkanian, s.f., p. 195; Fouqueré e Isaac, 2003; Deronne, 2011, p. 28). Hay quienes los contemplan incluso como los corpus en sí mismos (Kilgarriff y Grefenstette, 2003, citado por Tanguy, 2013, p. 2).

Sin embargo, paradójicamente, los colosales volúmenes de información de la Web, entorpecen y prolongan periódicamente la escogencia de las fuentes de acuerdo a necesidades, intereses y criterios concretos (Mehler *et al.*, 2010, p. 4). Se genera entonces un resultado contraproducente aun cuando se afine, se delimite o se filtre la búsqueda de las informaciones con las opciones avanzadas de motores como Google (Deronne, 2011, p. 43; Wissner, 2012, p. 246).

Para hablar de esto, algunos autores introducen la noción de «infobasura» (Hundt, Nesselhauf y Biewer, 2007, p. 69). Esta abarca, en cierto modo, la información fugaz, voluble (Fouqueré e Isaac, 2003), reiterativa, descontrolada, inútil, banal, pobre, con intereses ocultos o sin patrones homogéneos que simplifiquen su catalogación (Serres, 2005, p. 2-3). Internet no sería, por tal motivo, un corpus legítimo para muchos (Tanguy, 2013, p. 16).

La complejidad creciente radica en que ahora, cualquier información en Internet, es publicada sin que para ello sea imprescindible ser un profesional o un experto en una temática particular. En muchos casos, no se tienen en cuenta antiguos parámetros sobre la buena escritura, la ortografía, la redacción, la relectura, la edición o las reglas de publicación (Dagiral y Parasie, 2010; Simonnot, 2007; Côté y Troudi, s.f., p. 2; Mitou, 2006).

La calidad de las informaciones o la estructuración en las que deberían ser exhibidas pasan a menudo a un segundo plano y se prioriza la velocidad de visualización (Mehler *et al.*, 2010, p. 6; Dagiral y Parasie, 2010). En consecuencia, se torna arduo distinguir y catalogar los documentos de Internet, su exactitud y pertinencia (Booth, Colomb y Williams, 2008, p. 75). En últimas, sólo queda guiarse intuitiva y subjetivamente (Mehler *et al.*, 2010, p. 4) por el sentido común y la propia lógica (Serres, 2005, p. 2).

Los documentos que han seguido un estricto proceso de control y de corroboración constituyen las denominadas fuentes fiables. Ilustración de estas son determinados periódicos, libros, programas radiales o televisivos, artículos de revistas de divulgación, científicas o profesionales, documentos sobre congresos, datos estadísticos, vídeos, informes, ensayos, entre otros (Universidad de Alicante, s.f., p. 2). Usualmente, los autores de las fuentes fiables están adscritos a empresas, instituciones o entidades gubernamentales, educativas o investigativas reputadas. Los dominios de las fuentes fiables en Internet poseen abreviaturas tales como .edu o .gob, etc.

En lo que atañe a las fuentes no fiables, estas son las que mayoritariamente no han pasado por las etapas de revisión ni de edición tradicionales. Tampoco han sido publicadas por profesionales o expertos especializados, sino más bien, por gente común y corriente que regularmente hace primar una difusión digital apresurada.

Entre las fuentes no fiables se hallan los blogs, las páginas de periodismo ciudadano, las páginas personales, los foros o las páginas Wikipedia. No obstante, vale la pena aclarar que algunos de los contenidos que se encuentran en los ejemplos de fuentes fiables nombrados antes pueden ser no fiables (ej. Revistas no fiables, periódicos no fiables, etc.) (Sabrio y Burchfield, 2009, p. 225-227).

La palabra «blog» es el acrónimo de «weblog» (Lenormand, 2007, p. 91; Morand, Chevillat, Hrastnik y Jdey, 2006, p. 47) y alude a un periódico en línea inventado por un bloguero quien opina, comparte y recibe comentarios a propósito de temáticas particulares (Cobo, 2012, p. 27).

En ocasiones, las informaciones contenidas en los blogs no son nuevas sino que se repiten y se retoman de otros (Martínez y Solano, 2010, p. 87-88). Una variedad de

blogs son las llamadas páginas de periodismo ciudadano (Burgueño, 2010, p. 47) escritas con frecuencia por individuos sin formación en comunicación ni diseño editorial que cuentan noticias o que reflexionan sobre hechos de actualidad.

Las páginas personales y los foros exteriorizan opiniones, vivencias o puntos de vista sin garantía de objetividad (Larsonneur, 2008, p. 52).

Por último, Wikipedia (término proveniente de la expresión hawaiana «wiki wiki» que significa «rápido» (Martínez-Priego, 2012, p. 152), compuesto también de la partícula «pedia» que quiere decir «conocimiento»), se refiere a las páginas web colaborativas en donde los internautas redactan, complementan o corrigen trabajos que otros han escrito previamente.

Derivados de Wikipedia son la Wikimedia, el Wikcionario, la Wikiquote, la Wikiversidad, la Wikinoticias, la Wikisource o la Wikiespecies (Saorín, 2013). Todas tienen como primera meta divulgar gratuita y aceleradamente el conocimiento (Martínez, 2012, p. 10).

Por lo anterior, por su carácter de libre acceso participativo y así sea ampliamente consultada a escala mundial, es valorada como una fuente no fiable (Moya del amor, 2016, p. 11).

La tabla 1 muestra una lista de criterios a tener en cuenta a la hora de elegir y diferenciar las fuentes fiables de las no fiables.

Tabla 1. Criterios de identificación de las fuentes fiables y no fiables

Criterio	Explicación
Autoridad	Persona, entidad creadora del documento (Fornas, 2003, p. 76) o del sitio Internet (Maglione y Varlotta, s.f., p. 18-19). El renombre del autor, el peso de sus otras publicaciones y las citaciones de este, en otras fuentes (Bibliothèque de l'Université de Laval, 2011), se incluyen en esta pauta.
Credenciales, cualificación	Formación académica, experiencia, reconocimientos profesionales y científicos de los autores (Fornas, 2003, p. 76; Bibliothèque de l'Université de Laval, 2011; Nazario, Borchers y Lewis, 2010, p. 421).
Exactitud	Precisión de la información y adecuación con la realidad (Simonnot, 2007).
Inteligibilidad	Claridad, redacción, ortografía, estilo (Fornas, 2003, p. 76) y estructura del texto (Serres, 2005, p. 3).

Criterio	Explicación
Calidad y pertenencia	Ratificación de los contenidos con otras fuentes notorias (Simonnot, 2007; Maglione y Varlotta, s.f., p. 19), comprobación de plagio (Fornas, 2003, p. 80), revisión de comité evaluador, evaluación editorial (Serres, 2005, p. 3; Habert *et al.*, 1997, citado por Isaac *et al.*, s.f., p. 199), publicación en editorial prestigiosa (Booth *et al.*, 2008, p. 77).
Actualización	Puesta al día de la información (Maglione y Varlotta, s.f., p. 19).
Navegabilidad	Facilidad de uso del recurso o página (Fornas, 2003, p. 78; Universidad de Alicante, s.f., p. 5; Maglione y Varlotta, s.f., p. 20).
Independencia	Neutralidad e imparcialidad referente a intereses económicos, publicitarios o ideológicos (Fornas, 2003, p. 78; Bibliothèque de l'Université de Laval, 2011; Booth *et al.*, 2008, p. 77; Barker y Barker, 2014, p. 19).

METODOLOGÍA

Delimitación de la muestra

Para responder a las preguntas que formulamos en la introducción (a saber, si para la elaboración del corpus del proyecto de diccionario cultural contamos trabajar con fuentes fiables o análogamente con fuentes no fiables), principiamos escogiendo una muestra de 19 palabras o expresiones en francés. Tal vocabulario es portador de probables divergencias lingüísticas y culturales, totales o parciales, con respecto al español de Colombia y a la cultura de este país. La tabla 2 introduce el grupo de palabras o expresiones en francés.

Tabla 2. Palabras o expresiones de la muestra

Palabra o expresión francesa	Equivalente en español de Colombia
-Animal de compagnie	-Mascota
-Barbier	-Barbero
-Bureau de tabac	-Cigarrería
-Clavier	-Teclado

Palabra o expresión francesa	Equivalente en español de Colombia
-Concierge	-Conserje
-Dents du bonheur	-Dientes separados
-Écriture manuscrite	-Escritura manuscrita
-Gant de toilette	-Guante de baño
-Guinguette	-Especie de bar restaurante situado cerca de una fuente de agua (ej. Río, lago, etc.)
-Heure d'été	-Hora de verano
-Klaxon	-Pito
-Marchander	-Regatear
-Œillet	-Clavel
-Pantalon	-Pantalón
-Quatre-vingt(s)	-Ochenta
-Restaurant	-Restaurante
-Sèyes	-Sèyes
-Strapontin	-Silla plegable
-Stylo-plume	-Pluma

Colecta de informaciones

Una vez las palabras y expresiones seleccionadas, para este estudio, procedimos a formar dos corpus en francés. El primero se compone de fuentes fiables y el segundo de fuentes no fiables. Los dos, se fundamentan en informaciones extraídas de Internet sobre el léxico escrutado.

Con el objetivo de recolectar los datos para cada una de las palabras o expresiones, nos servimos de Google (2017). A través de la herramienta de exploración avanzada, filtramos las páginas web de Francia y localizamos enlaces de interés digitando

palabras o expresiones claves sobre la muestra. Los resultados, hasta el número veinte de cada búsqueda en Google, fueron tomados en consideración.

Para agrupar los enlaces, fruto de nuestra indagación, en el corpus de fuentes fiables o en el corpus de fuentes no fiables, establecimos una clasificación de los sitios o páginas Internet de acuerdo a los conceptos examinados en el marco teórico. La tabla 3 expone dicha catalogación.

Tabla 3. Clasificación de documentos en fuentes fiables y no fiables

Páginas o sitios fiables	**Páginas o sitios no fiables**
-Artículos de periódicos en línea reconocidos	-Páginas Wiki
-Artículos de revistas especializadas	-Blogs
-Sitios web especializados	-Páginas personales
-Páginas estatales o gubernamentales	-Foros
-Sitios comerciales especializados	-Sitios anónimos
-Blogs de periódicos y sitios distinguidos	-Revistas o periódicos poco reconocidos
-Libros en línea	-Páginas con intereses comerciales o ideológicos
-Páginas de noticias	-Diarios personales
-Páginas de entidades regionales o administrativas	-Páginas de periodismo participativo
-Páginas de asociaciones notables	-Páginas de anuncios de empleo de particulares
-Páginas de bibliotecas	-Páginas de informaciones y consejos realizadas por gente no conocida
-Páginas de informativos de radio y televisión	
-Páginas de informes	
-Páginas de institutos de sondeos y encuestas	
-Tesis	

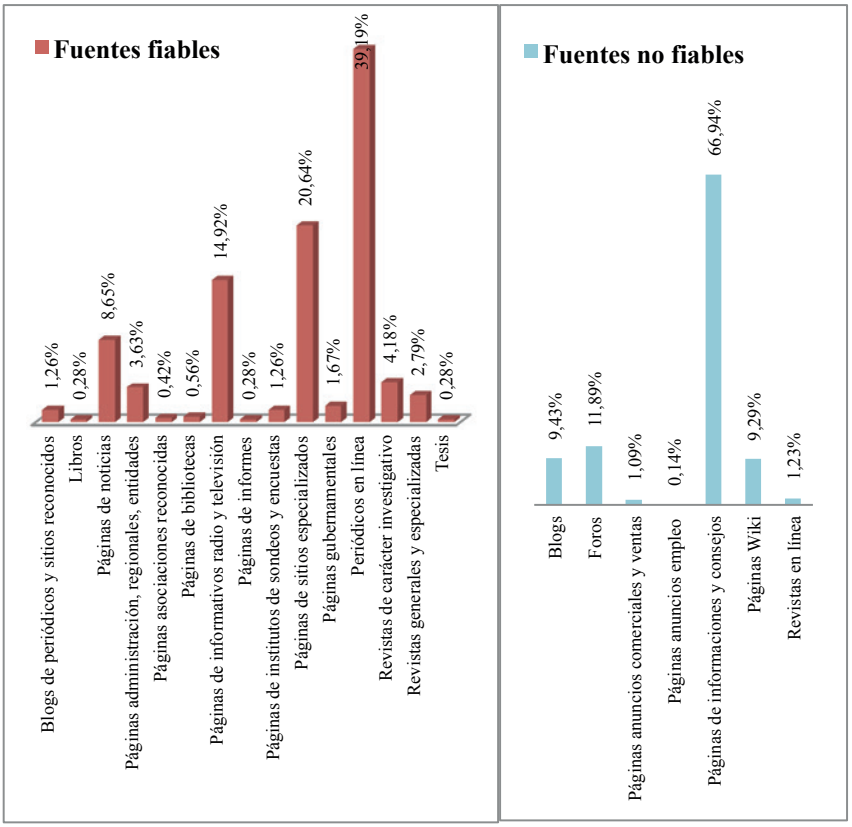
Los datos fueron registrados con la ayuda del programa de análisis de textos «Texstat» (Hüning, 2002). Esta herramienta, además de hacer viable la constitución de los corpus, genera una lista de frecuencias y establece concordancias de las palabras o expresiones para examen posterior.

Las concordancias, al tratarse de ejemplos en donde está la palabra o expresión clave en contexto, posibilitan la identificación de informaciones lingüísticas y culturales importantes. Los análisis de las concordancias son cuantitativos ya que se suministra un número dado de las mismas. De manera similar, son cualitativos porque se estudia el texto que va antes y después de la palabra o expresión clave (Gatto, 2014, p. 24).

ANÁLISIS DE LOS DATOS Y RESULTADOS

Según datos compilados, 732 páginas (con un total de 99.803 palabras) proceden de fuentes no fiables y 717 páginas (con un total de 83.356 palabras) pertenecen a fuentes fiables. La figura 1 resume los géneros o tipos de páginas web con sus respectivos porcentajes de aparición que fueron concatenados en el corpus fiable y en el no fiable.

Figura 1. Tipos de fuentes digitales con porcentaje de aparición

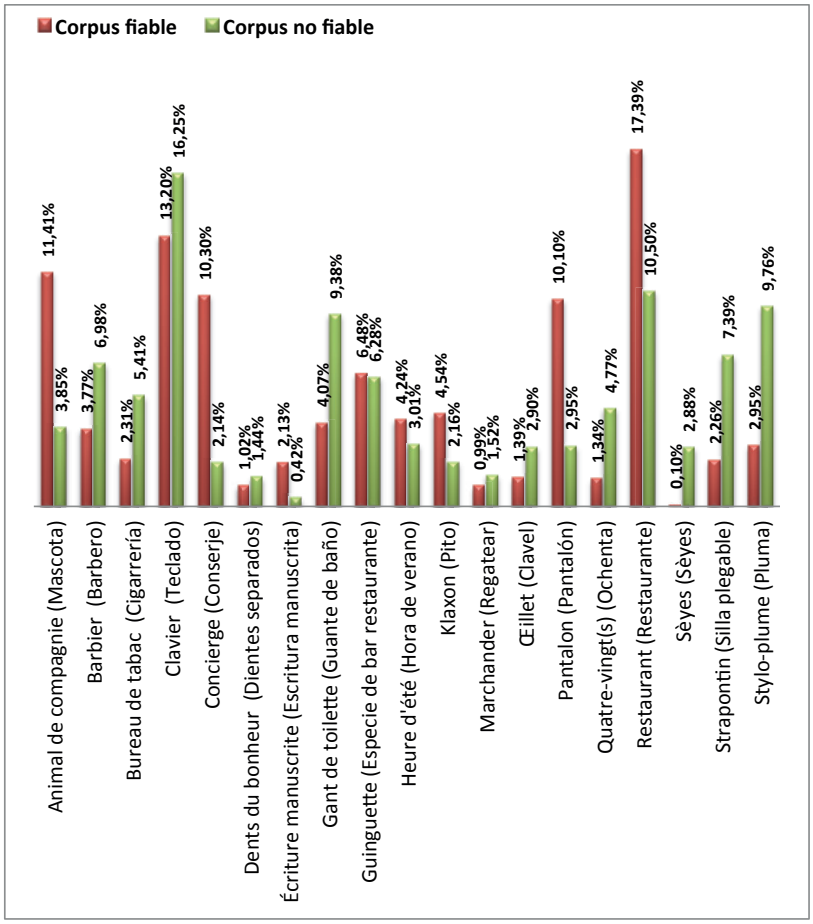


Entre las páginas fiables, fue más común dar con informaciones relativas a nuestra muestra de palabras, en los periódicos en línea (39,19 %), seguidos de las páginas de sitios especializados (20,64 %), de las páginas de informativos radiales y televisivos (14,92 %) y de otras páginas de noticias (8,65 %).

En lo que atañe al corpus de fuentes no fiables, los géneros que presentaron mayor número de información sobre las palabras y expresiones de la muestra fueron las páginas de informaciones y consejos (66,94 %), los foros (11,89 %), los blogs (9,43 %) y las páginas Wiki (9,29 %).

Enfocándonos en las palabras y expresiones tenidas en cuenta para este estudio preliminar, 11 de las 19 poseen frecuencias más elevadas en el corpus no fiable y 8 de 19 se cuentan más veces en el corpus fiable. Esto se observa en la figura 2.

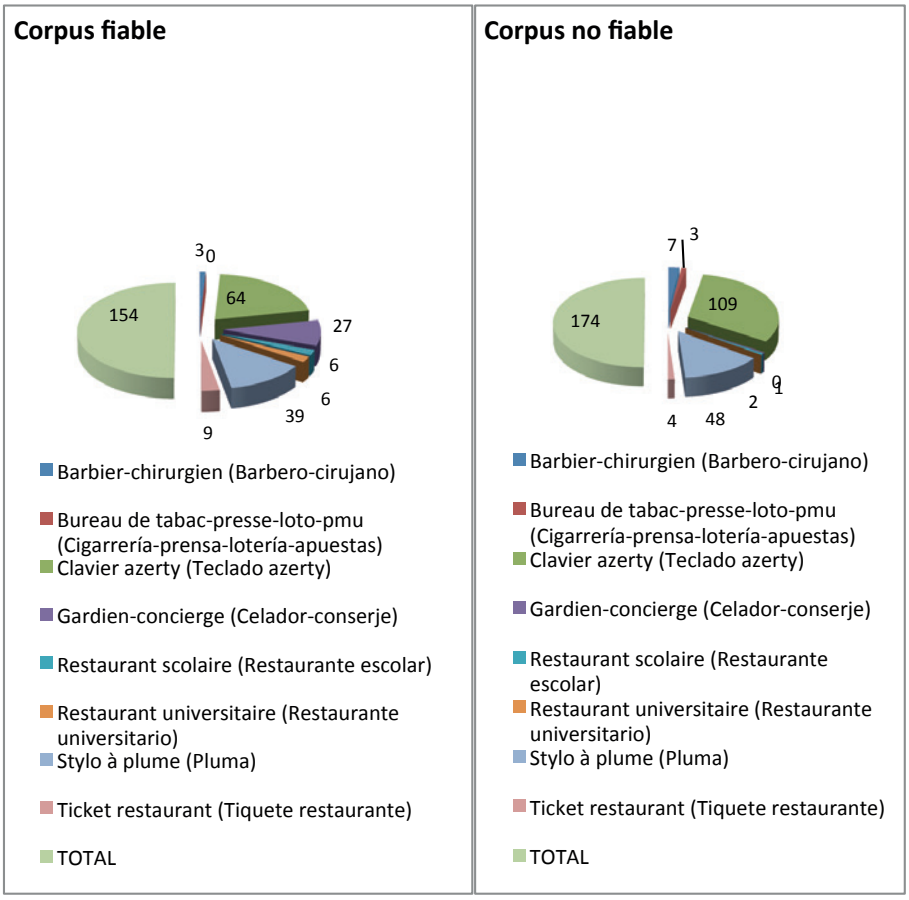
Figura 2. Porcentajes de aparición de las palabras o expresiones en cada uno de los corpus



Ahora bien, con miras a redactar los artículos de diccionario cultural para nuestro proyecto, en ambos corpus se contemplan informaciones lingüísticas y culturales. Mencionamos las ligadas a formas de ver el mundo y de estructurar la realidad, a actitudes, comportamientos, creencias, rituales, tradiciones, a sintagmas complejos sobre las palabras o expresiones claves, a categorías gramaticales, glosas definicionales, explicativas o descriptivas, a palabras en contexto, sinónimos, antónimos, variantes de escritura, familias de palabras, informaciones etimológicas, entre otros.

A manera de ejemplo, ilustramos en la figura 3 sintagmas complejos relacionados con los términos de la muestra, con su número de ocurrencias en ambos corpus.

Figura 3. Ejemplos de sintagmas complejos sobre términos de la muestra



El diagrama anterior señala la frecuencia total de las expresiones tomadas en consideración, tanto en el corpus fiable, como en el corpus no fiable. Se constata que el corpus no fiable brinda un mayor número de frecuencias del conjunto de expresiones nombradas (174 contra 154 para el corpus fiable).

En los dos corpus, la mayoría de los sintagmas aparecen, lo que ratifica las informaciones dispensadas por el corpus no fiable con respecto al corpus fiable.

Los dos corpus son del mismo modo complementarios. Algunas de las expresiones que no estaban en el corpus de fuentes fiables se hallan en el corpus de fuentes no fiables y viceversa (ej. «Bureau de tabac-presse-loto-pmu» (cigarrería-prensa-lotería-apuestas) o «restaurant universitaire» (restaurante universitario)).

Esto se corrobora similarmente en la tabla 4 con los hápax de posibles expresiones. Aunque se encuentran sólo una vez en uno de los dos corpus, podrían ser significativos al redactar nuestros artículos de diccionario, tras verificación de su existencia y utilización.

Tabla 4. Ejemplos de hápax en fuentes fiables y no fiables

Hápax con equivalente en español	**Corpus fiable**	**Corpus no fiable**
Pantalon à pinces (Pantalón de prenses)	1	0
Pantalon à taille base (Pantalón de talle bajo)	0	1
Pantalon à taille haute (Pantalón de talle alto)	0	1
Pantalon moulant (Pantalón apretado)	1	1
Restaurant d'entreprise (Restaurante de empresa)	1	1
Tabac-presse-jeu (Cigarrería-prensa-juego)	0	1
Total	3	5

A modo ilustrativo, en las figuras 4 a 6, contraponemos igualmente las informaciones ubicadas en ambos corpus sobre la palabra «quatre-vingt(s)» (ochenta) y las expresiones «coup de klaxon» (pitazo) y «Nouveaux animaux de compagnie» (nuevas

mascotas). Nos percatamos que, en ocasiones, el corpus fiable y el corpus no fiable brindan datos lingüísticos y culturales análogos y complementarios.

En las gráficas de las figuras 4, 5 y 6, el intitulado «otros» alude al número de informaciones infructuosas que no aportan datos lingüísticos ni culturales relevantes para la redacción de los artículos de diccionario cultural. Vemos que estos datos inservibles están presentes tanto en el corpus fiable como en el corpus no fiable.

Figura 4. Datos lingüísticos y culturales de la palabra «quatre-vingt(s)» (ochenta)

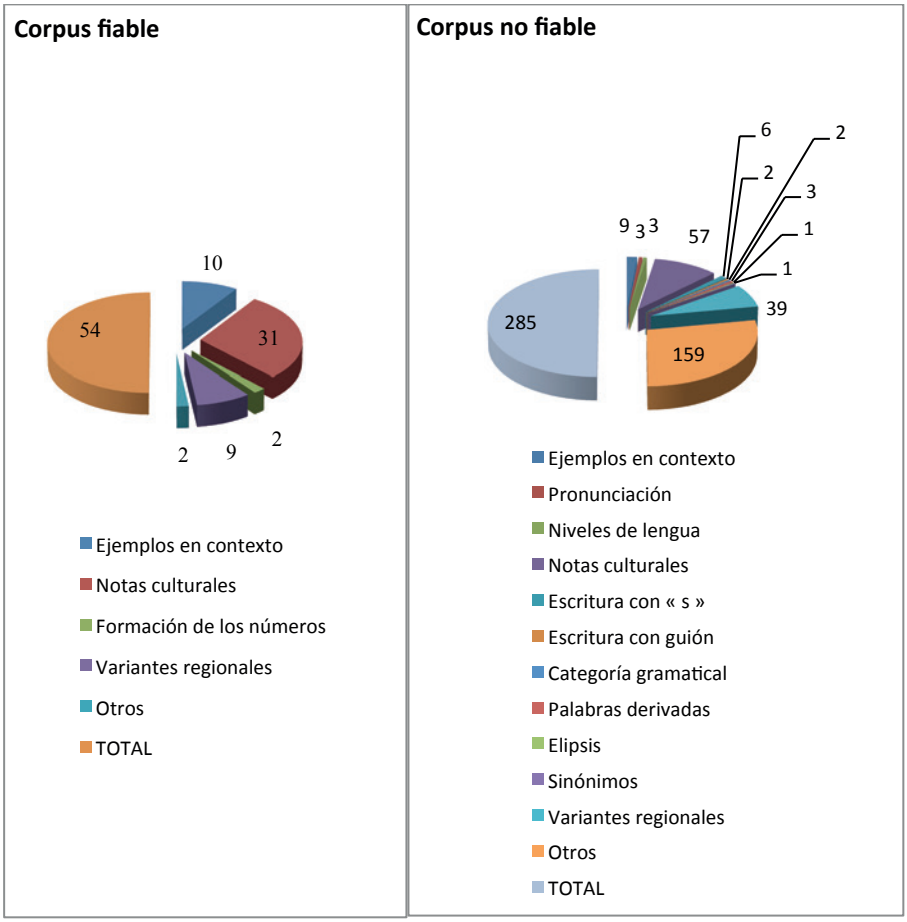


Figura 5. Datos lingüísticos y culturales de la expresión «coup de klaxon» (pitazo)

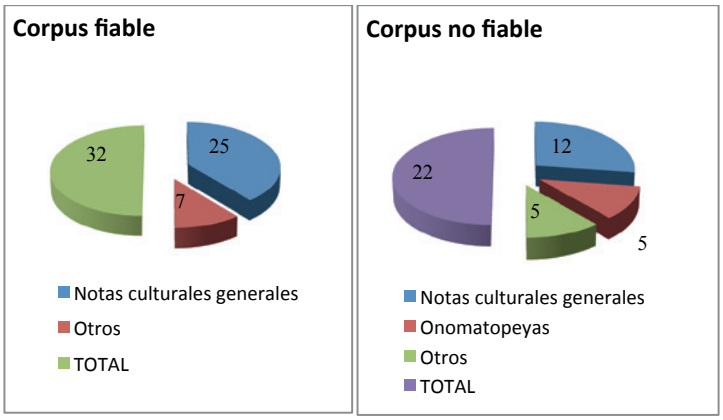
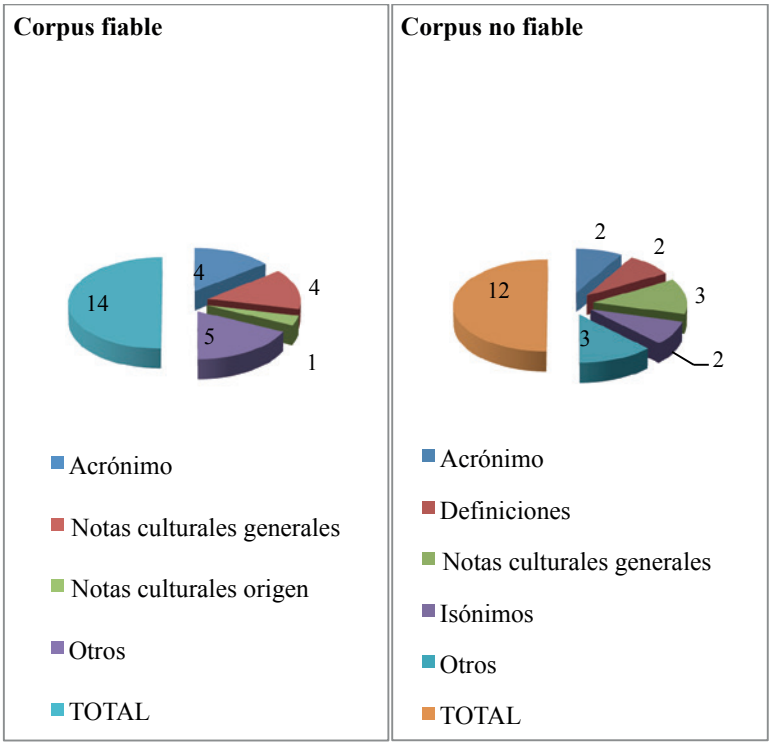


Figura 6. Datos lingüísticos y culturales de la expresión «Nouveaux animaux de compagnie» (nuevas mascotas)



Tras haber esquematizado casos cuantitativos concretos, enseguida, ponemos de relieve extractos comparativos con informaciones lingüísticas y culturales de algunas de las palabras y expresiones de la muestra por cada uno de los corpus. A través de estos, se confirma que los contenidos que están en las fuentes fiables y en las no fiables vehiculan muchas veces la misma información.

Esto se manifiesta en la tabla 5 con aquellos que detallan notas culturales y variantes regionales («huitante») del número «ochenta» («quatre-vingt(s)»); el sinónimo («dents de la chance») de la expresión «dientes separados» («dents du bonheur»); informaciones sobre el nombre de marca («klaxon») de la palabra «pito»; variantes de escritura («stylo plume») de la palabra «pluma» («stylo-plume»); el acrónimo («NAC») de la expresión «nuevas mascotas» («Nouveaux animaux de compagnie») o el léxico relativo a los niveles de lengua («resto») de la palabra «restaurante» («restaurant»).

Tabla 5. Ejemplos de informaciones lingüísticas y culturales similares en ambos corpus

Corpus de fuentes fiables	**Corpus de fuentes no fiables**
Quatre-vingt(s) (Ochenta)	
«qui était un système vicésimal, (...). Ce système était courant chez les peuples d'origine germanique. Selon ce système, on trouvait les formes (...) quatre vins pour 80» (Bibliothèque municipale de Lyon, 2016). «que era un sistema vigesimal, (...). Este sistema era común en los pueblos de origen germánico. Según este sistema, se encontraban las formas (...) cuatro veintes para 80» (traducción de Bibliothèque municipale de Lyon, 2016).	«Quatre-vingts (80) provient probablement de l'ancien système vigésimal de numération (4×20)» (Wiktionary, 2017a). «Ochenta (80) proviene probablemente del antiguo sistema de numeración vigesimal (4x20)» (traducción de Wiktionary, 2017a).
«pourquoi la Suisse et la Belgique disent huitante, nonante et pas la France ?» (Bibliothèque municipale de Lyon, 2016). «¿Por qué Suiza y Bélgica dicen «huitante», «nonante» y no Francia?» (traducción de Bibliothèque municipale de Lyon, 2016).	«nous devons nous convertir, dès aujourd'hui, au Septante, huitante, et Nonante ...Comme le font nos amis Belges et Suisses» (Forum France 2, 2006). «Debemos convertirnos de ahora en adelante al «septante», «huitante» et «nonante»... Como lo hacen nuestros amigos belgas y suizos» (traducción de Forum France 2, 2006).

Corpus de fuentes fiables	**Corpus de fuentes no fiables**
Dents du bonheur (Dientes separados)	
«nous appelons cela «les dents de la chance» et plus souvent encore «les dents du bonheur»» (Arte-Karambolage, 2015). «llamamos esto «los dientes de la suerte» y más comúnmente «los dientes de la felicidad»» (traducción de Arte-Karambolage, 2015).	«on parle parfois de «dents du bonheur» ou «dents de la chance»» (Culture-générale.fr, s.f.). «se habla, en ocasiones, de «dientes de la felicidad» o «dientes de la suerte»» (traducción de Culture-générale.fr, s.f.).
Klaxon (Pito)	
«Klaxon est une marque déposée par la société du même nom» (Le Figaro.fr, 2007). «Klaxon es una marca registrada por la sociedad que posee el mismo nombre» (traducción de Le Figaro.fr, 2007).	«Le mot «klaxon» (...) C'est une marque commerciale déposée par la société Klaxon Signals Ltd» (TeleCarteGrise, 2009). «La palabra «klaxon» (...). Es una marca comercial registrada por la sociedad Klaxon Signals Ltd» (traducción de TeleCarteGrise, 2009).
Stylo-plume (Pluma)	
«le stylo plume et tout le cérémonial» (Le Monde.fr, 2015). «la pluma y todo el ceremonial» (traducción de Le Monde.fr, 2015).	«je nettoyais mes stylos plumes avec de l'eau» (Forum Doctissimo, 2008). «limpiaba mis plumas con agua» (traducción de Forum Doctissimo, 2008).
Animaux de compagnie (Mascotas)	
«La tendance des NAC se confirme d'une année sur l'autre» (La Dépêche.fr, 2014). «La tendencia de las nuevas mascotas se confirma de un año al otro» (traducción de La Dépêche.fr, 2014).	«Les nouveaux animaux de compagnie (plus généralement nommés par l'acronyme NAC)» (Wikipédia, 2013). «Las nuevas mascotas (más conocidas por el acrónimo «NAC»)» (traducción de Wikipédia, 2013).
Restaurant (Restaurante)	
«elle propose bien plus qu'un resto dansant» (Le Figaro.fr, 2011). «propone más que un restaurante bailable» (traducción de Le Figaro.fr, 2011).	«Un resto en solo, une pratique de plus en plus répandue» (Snacking, 2017). «Un restaurante en solitario, una práctica cada vez más generalizada» (traducción de Snacking, 2017).

En otras situaciones, informaciones peculiares aparecen en uno u otro corpus pero no en los dos. Ejemplificamos esto en la tabla 6 con las de tipo etimológico sobre la palabra «silla plegable» («strapontin»), con el nombre científico («dianthus») de la palabra «clavel» («œillet»), con la nota cultural de la palabra «dientes separados» («dents du bonheur») o con la nota lingüística sobre la palabra «ochenta» («quatre-vingt(s)») que figuran solamente en el corpus de fuentes no fiables.

Tabla 6. Ejemplos de informaciones lingüísticas y culturales encontradas únicamente en el corpus de fuentes no fiables

Corpus de fuentes fiables	**Corpus de fuentes no fiables**
Strapontin (Silla plegable)	
No se encontraron informaciones etimológicas.	«De l'italien strapuntino («matelas de marin, hamac»)» (Wiktionary, 2017b). «Del italiano strapuntino («colchón de marino, hamaca»)» (traducción de Wiktionary, 2017b).
Œillet (Clavel)	
No se encontraron informaciones sobre el nombre científico.	«Connu sous son nom latin «dianthus»» (Prestigemaison.com, 2017). «Conocido bajo el nombre científico de «dianthus»» (traducción de Prestigemaison.com, 2017).
Dents du bonheur (Dientes separados)	
No se encontraron informaciones sobre el origen de la expresión.	«cette expression tire son origine du temps de Napoléon ! (...) les soldats aient des incisives en parfait état, car ils devaient ouvrir leur poudrière avec les dents afin de recharger leur fusil qu'ils devaient tenir à deux mains. Tous ceux qui avaient des dents écartées étaient alors réformés car inaptes au combat... et ce, pour leur plus grand bonheur !» (Ça m'intéresse, 2010). «esta expresión proviene de la época de Napoleón (...) que los soldados tengan incisivos en perfecto estado, porque tenían que abrir el polvorín con los dientes para volver a cargar el fusil que sujetaban con las dos manos. Todos los que tenían los dientes separados eran excluidos porque eran inaptos para el combate... ¡Y esto provocaba gran felicidad!» (traducción de Ça m'intéresse, 2010).

Corpus de fuentes fiables	Corpus de fuentes no fiables
Quatre-vingt(s) (Ochenta)	
No se encontraron informaciones lingüísticas sobre el guión.	«S'écrit toujours avec un trait d'union. Quatre-vingts hommes» (Wiktionary, 2017a). «Se escribe siempre con un guión. Ochenta hombres» (traducción de Wiktionary, 2017a).

Finalmente, en la tabla 7, ilustramos informaciones que sólo están en el corpus de fuentes fiables. Es el caso de la expresión «pantalón rojo vivo» («pantalon garance») que se acompaña de una nota cultural.

Tabla 7. Ejemplo de información cultural encontrada únicamente en el corpus de fuentes fiables

Corpus de fuentes fiables	Corpus de fuentes no fiables
Pantalon garance (Pantalón rojo vivo)	
«la dramatique histoire du pantalon garance. Dès le début de la guerre 1914-1918, des milliers de soldats français ont été fauchés, victimes de la couleur rouge vif de leurs pantalons» (Bessard, 2014). «la dramática historia del pantalón rojo vivo. Desde comienzos de la guerra de 1914-1918, miles de soldados franceses fueron asesinados, víctimas del color rojo vivo de sus pantalones» (traducción de Bessard, 2014).	No se encontraron informaciones sobre esta expresión.

CONCLUSIONES

La preferencia de determinadas fuentes, en detrimento de otras, está sujeta a los objetivos que anhela alcanzar el lector o el investigador. Hay quienes se regirán concretamente por fuentes fiables. Otros, se interesarán más bien en las fuentes no fiables porque quizás ejecutan proyectos que indagan sobre fenómenos lingüísticos, sociales o culturales particulares.

En nuestro proyecto de diccionario cultural, al ser un trabajo investigativo y educativo, concebimos, en primera instancia, enfatizarnos en el uso de fuentes fiables. Sin embargo, los datos conseguidos en este estudio nos permiten afirmar que no vamos

a descartar las fuentes no fiables puesto que ofrecen informaciones lingüísticas y culturales suplementarias que ahondan los datos a tener en consideración para la investigación.

Los foros, los blogs las páginas personales, las Wiki y otras fuentes no fiables son muchas veces documentos auténticos escritos por hablantes nativos de la lengua y de la cultura francesa o colombiana. Por esta razón, proveen ejemplos representativos, reales, implícitos o explícitos, sobre hechos lingüísticos y culturales u otros elementos que tal vez no están en las fuentes fiables.

Del mismo modo, hay que agregar que pese a que los datos no estén en fuentes fiables y no hayan sido escritos por especialistas o expertos, no supone necesariamente que sean de mala calidad (Noël, 2007, p. 67). Por consiguiente, a pesar de la prioridad atribuida a las fuentes fiables, es posible servirse asimismo de fuentes no fiables. En este último caso, se deben contrastar o comparar con las primeras para garantizar la veracidad de las informaciones.

Por otro lado, la enorme cantidad de referencias de Internet hace laborioso distinguir las fuentes confiables. Por momentos, es esencial orientarse por la lógica y la intuición. La clasificación de las fuentes en fiables y no fiables se convierte de esta manera en algo subjetivo.

Para este análisis, dudamos en clasificar ciertas fuentes en fiables o en no fiables. Esto ocurrió con páginas de noticias, sitios comerciales, tesis, revistas o blogs. En efecto, al no ser de autores renombrados, nos interrogamos sobre su calidad. Similarmente, controvertimos sobre el hecho de que sitios específicos, aunque tuvieran intereses económicos, brindaban informaciones fructíferas. Otros, aun cuando departieran de temas un tanto banales podían haber pasado por procesos de calidad exigentes.

Se concluye, análogamente, que la forma en la que realicemos la búsqueda influye en el número de enlaces hallados y en los resultados.

REFERENCIAS

Arte-Karambolage. (2015). *Le mail : Les dents du bonheur*. Recuperado de http://sites.arte.tv/karambolage/fr/le-mail-les-dents-du-bonheur-karambolage

Austermühl, F. (2014). *Translation Practices Explained*. Oxon: Routledge.

Barker, D. y Baker, M. (2014). *Internet Research*. Stamford: Cengage Learning.

Bessard, M. (2014). *Péronne : La dramatique histoire du pantalon garance*. Recuperado de http://france3-regions.francetvinfo.fr/bourgogne-franche-comte/2014/08/30/peronne-la-dramatique-histoire-du-pantalon-garance-540248.html

Bibliothèque de l'Université de Laval. (2011). *Info Sphère. Évaluer la qualité des sources*. Recuperado de https://www.bibl.ulaval.ca/infosphere/sciences_humaines/evaeva1.html

Bibliothèque municipale de Lyon. (2016). *Nonante*. Recuperado de http://www.guichetdusavoir.org/viewtopic.php?f=2&t=67321&p=129474&hilit=hommes&sid=5b56287a6c437b88e408f7e0e88a5d17

Booth, W., Colomb, G. y Williams, J. (2008). *The Craft of Research*. Chicago: The University of Chicago Press.

Burgueño, J. (2010). *Cuestión de confianza. La credibilidad, el último reducto del periodismo del siglo XXI*. Barcelona: Editorial UOC.

Ça m'intéresse. (2010). *D'où vient l'expression «avoir les dents du bonheur» ? Overblog.* Recuperado de http://ca-m-interesse.over-blog.com/article-d-ou-vient-l-expression-avoir-les-dents-du-bonheur-49540602.html

Clarenc, C. (2011). *Nociones de cibercultura y periodismo*. S.l.: Lulu.com.

Cobo, S. (2012). *Internet para periodistas. Kit de supervivencia para la era digital*. Barcelona: Editorial UOC.

Côté, M. y Troudi, N. (S.f.). *NetSA : Une achitecture multiagent pour la recherche sur Internet*. Recuperado de http://www2.ift.ulaval.ca/~chaib/publications/cote.pdf

Culture-générale.fr. (S.f.). *Qu'est-ce qu'un diastème ?* Recuperado de https://www.culture-generale.fr/expressions/4840-quest-ce-quun-diasteme

Dagiral, É. y Parasie, S. (2010). Presse en ligne : Où en est la recherche ? *Réseaux, 2*(160-161), 13-42. doi: 10.3917/res.160.0013

Deronne, E. (2011). Intérêt et pièges de la Toile en tant que corpus pour la recherche en linguistique (sous l'angle de recherches sur la valence verbale). *Revue Tranel, 55,* 25-44.

Drago, L. (2009). Le Web comme corpus : Documents authentiques et exploitation en FLE. *Mélanges CRAPEL*, (31), 237-257.

Flowerdew, L. (2012). *Corpora and Language Education*. Great Britain: Springer.

Fornas, R. (2003). Criterios para evaluar la calidad y fiabilidad de los contenidos en Internet. *Rev. Esp. Doc. Cient., 26*(1), 75-80.

Forum Doctissimo. (2008). *Changer de couleur l'encre de mon stylo plume ?* Recuperado de http://forum.doctissimo.fr/viepratique/Astuces-maison-et-linge/changer-couleur-plume-sujet_1370_1.htm

Forum France 2. (2006). *Sujet : Soixante-dix, quatre-vingt-dix = Royalistes ?* Recuperado de http://forums.france2.fr/france2/onnapastoutdit/soixante-quatre-royalistes-sujet_22589_1.htm

Fouqueré, C. e Isaac F. (2003). Corpus issus du Web : Constitution et analyse informationnelle. *Revue québécoise de linguistique, 32*, 111-134.

Gatto, M. (2014). *Web as corpus. Theory and practice. Studies in corpus and discourse*. New York: Bloomsbury.

Google. (2017). *Google*. Recuperado de https://www.google.fr/

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

Hüning, M. (2002). *Textstat*. Recuperado de http://neon.niederlandistik.fu-berlin.de/en/textstat/

Isaac, F., Hamon, T., Fouqueré, C., Bouchard, L. y Emirkanian, L. (S.f.). *Extraction informatique de données sur le web*. Recuperado de http://cqfd.teluq.uquebec.ca/distances/D5_2_o.pdf

La Dépêche.fr. (2014). *Animaux de compagnie : La France aux 63 millions d'amis…* Recuperado de http://www.ladepeche.fr/article/2014/11/16/1992325-animaux-de-compagnie-la-france-aux-63-millions-d-amis.html

Larsonneur, C. (2008). *La recherche Internet en lettres et langues*. Paris: Éditions Ophrys.

Le Figaro.fr. (2007). *Klaxon, Kleenex, Rimmel, Mobylette, Frigidaire… Ces noms propres devenus communs*. Recuperado de http://www.lefigaro.fr/economie/2007/03/15/04001-20070315ARTFIG90103-klaxon_kleenex_rimmel_mobylette_frigidaire_ces_noms_propres_devenus_communs.php

Le Figaro.fr. (2011). *Nogent, Joinville, Bry, Champigny, la douceur de vivre sur les bords de Marne.* Recuperado de http://www.lefigaro.fr/sortir-paris/2011/06/21/03013-20110621ARTFIG00580-nogent-joinville-bry-champigny-la-douceur-de-vivre-sur-les-bords-de-marne.php

Le Monde.fr. (2015). *Plumes aériennes*. Recuperado de http://www.lemonde.fr/m-le-mag/article/2015/03/10/plumes-aeriennes_4587758_4500055.html

Lenormand, P. (2007). *Internet : Techniques de recherche pour les professionnels*. Paris: Éditions Eni.

Maglione, C. y Varlotta, N. (S.f.). *Investigación, gestión y búsqueda de información en Internet*. Recuperado de http://bibliotecadigital.educ.ar/uploads/contents/investigacion0.pdf

Martínez, C. (2012). *Wikipedia inteligencia colectiva en la red*. Barcelona: Profit editorial.

Martínez, S. y Solano, E. (2010). *Blogs, bloggers, blogósfera. Una revisión multidisciplinaria*. México: Universidad Iberoamericana.

Martínez-Priego, C. (2012). *Quiero ser Community Manager*. Madrid: ESIC Editorial.

McEnery, T., Xiao, R. y Tono, Y. (2006). *Corpus-Based Language Studies: an Advanced Resource Book*. Oxon: Routledge.

Mehler A., Sharoff, S. y Santini, M. (2010). *Genres on the Web. Computational Models and Empirical Studies*. Germany: Springer.

Mitou, D. (2006). *L'évaluation de l'information sur Internet*. Recuperado de http://bbf.enssib.fr/consulter/bbf-2006-05-0091-005

Morand, J.-C., Chevillat, J., Hrastnik, R. y Jdey, A. (2006). *RSS, blogs, un nouvel outil pour le management*. Paris: M2 Éditions.

Moya del amor, M. (2016). *Los ríos temporales en la red: Un análisis en la Wikipedia*. Alcoy: 3Ciencias.

Müller, A. y Gjerstad, Ø. (2014). Web 2.0 et genres discursifs : L'exemple de blogs sur le changement du climat. *Synergies Pays Scandinaves*, (9), 49-61.

Nazario, L., Borchers, D. y Lewis, W. (2010). *Bridges to Better Writing*. Boston: Wadsworth, Cengage Learning.

Noël, É. (2007). *Évaluer l'information sur Internet*. Recuperado de https://issuu.com/elisaformist/docs/brochurerepere

Olivier, A., Moré, J. y Climent, S. (2008). *Traducción y tecnologías*. Barcelona: Editorial UOC.

Prestigemaison.com. (2017). *Symbole de l'œillet : Fleur de naissance de janvier*. Recuperado de http://www.prestigemaison.com/symbole-de-loeillet-fleur-de-naissance-de-janvier/

Sabrio, D. y Burchfield, M. (2009). *Insightful Writing*. Boston: Cengage Learning.

Saorín, T. (2013). *Wikipedia, de la A a la W*. Barcelona: Editoral UOC.

Serres, A. (2005). Évaluation de l'information : Le défi de la formation. *Bulletin des bibliothèques de France, French School of Librarianship and Information Science*, (6), 1-10.

Simonnot, B. (2007). Évaluer l'information. *Documentaliste Sciences de l'information*, *3*(44), 210-216.

Snacking. (2017). *Un resto en solo, une pratique de plus en plus répandue*. Recuperado de https://www.snacking.fr/news-3344-Un-resto-en-solo--une-pratique-de-plus-en-plus-repandue---.php

Tanguy, L. (2013). La ruée linguistique vers le Web. *Texto ! Textes et Cultures, Équipe Sémantique des textes, 18*(4), 1-33.

TeleCarteGrise. (2009). *Accros du klaxon, prenez garde !* Recuperado de http://www.telecartegrise.com/flash-actu-carte-grise/2014/12/20141202-FILWWW000040-Accros-du-klaxon-prenez-garde.html

Terrádez, M. (2001). *Frecuencias léxicas del español coloquial: Análisis cuantitativo y cualitativo*. Valencia: Universidad de Valencia.

Universidad de Alicante. (S.f.). *Cómo evaluar la información encontrada*. Recuperado de https://rua.ua.es/dspace/bitstream/10045/46567/1/ci2_avanzado_2014-15_Como-evaluar-informacion.pdf

Wissner, I. (2012). Les grands corpus du français moderne : Des outils pour étudier le lexique diatopiquement marqué ? *Sky Journal of Linguistics, 25*, 233-272.