# Corpus Lingüísticos del Español[1]

**Rodolfo Bonino**
Universidad Nacional de Rosario
Facultad de Humanidades y Artes
rodolfobonino@yahoo.com.ar

## Resumen

En este trabajo se reseña el lugar que ocupan los datos en distintas corrientes lingüísticas, se explica cómo se ha constituido un corpus específico para el análisis computacional de las construcciones de *hacer* + infinitivo a partir del *Corpus de Referencia del Español Actual* (*CREA*) y cómo se los ha organizado en una base de datos relacional. Con el objeto de comprobar que el corpus específico se encuentra dentro de los parámetros generales y si resulta adecuado para la investigación, se establecen comparaciones cuantitativas de los datos de ese corpus con otros obtenidos de *El Grial* y del *Corpus del Español.* La comparación cualitativa orienta acerca de cuál es el recurso más eficaz para el tratamiento informático.

**Palabras claves:** Corpus lingüísticos del español. Construcciones de *hacer* + infinitivo. Propiedades cuantitativas. Tratamiento informático.

## Abstract

In this work, a description is provided on the role that data occupy in different linguistic trends, explanation is given of the way in which a specific corpus was built for the computerized analysis of the *hacer + INFINITIVO* construction from the *Corpus de Referencia del Español Actual* (*CREA*), and the way in which these data have been arranged in a network relation database. Quantitative comparisons between data from this corpus and from *El Grial* and the *Corpus del Español* are made in order to check whether the specific corpus falls within general parameters and whether it proves suitable for research. Qualitative comparison shows which the most effective resource for their processing is.

**Key words:** Linguistic corpora on Spanish language. *Hacer + INFINITIVO* constructions. Data processing.

## 1. INTRODUCCIÓN

En la historia de los estudios del lenguaje el lugar de los datos ha ido variando según las diferentes concepciones teóricas, la metodología de investigación utilizada, los objetivos perseguidos y los recursos tecnológicos disponibles.

De distintas etapas de esta evolución histórica quedaron colecciones importantes de datos que muestran la impronta su origen teórico - metodológico. En lo que respecta a la lengua española, se

[1] Este trabajo forma parte de mi tesis de doctorado dirigida por la Dra. Zulema Solana: *Construcciones de* hacer + *infinitivo. Descripción, análisis y tratamiento informático.*

pueden citar el *Diccionario de Autoridades* de la Real Academia Española (1726 – 1739), el *Diccionario de Construcción y Régimen*[2] de Rufino Cuervo y en los *Archivo Gramatical de la Lengua Española (AGLE)* de Salvador Fernández Ramírez[3].

El objeto de estudio de la gramática tradicional no era el lenguaje en su totalidad sino solamente aquellas manifestaciones lingüísticas consideradas cultas; por lo tanto, los únicos datos relevantes eran los que provenían de autores consagrados y que podían ser tomados como ejemplares. En el caso del *Diccionario de Autoridades*, que es la obra más antigua, se puede inferir que, si bien el origen de los datos (principalmente, autores consagrados) tiene un papel preponderante para establecer el modelo de lengua culta, se aplica un segundo filtro selectivo, que es la "doctrina gramatical", cuyos principios, aunque no muy claros, se basan, sobre todo, en la lógica aristotélica y la etimología.

El trabajo de Cuervo parece dar más relevancia a los datos, pero continúa presentando una fuerte impronta de la tradición gramatical.

En cambio, aunque no se pueden incluir en una lingüística empírica, los archivos de Fernández Ramírez fueron realizados en el horizonte teórico de la lingüística del Siglo XX y el los datos constituyen el fundamento para un estudio sistemático de la lengua general.

Dado que hasta las últimas décadas el único recurso tecnológico disponible fue la imprenta, el ordenamiento alfabético de las entradas léxicas era el modo más sencillo y eficaz de acceso a los datos; de ahí que las primeras fuentes de datos fuesen los diccionarios. Las cédulas de Fernández Ramírez no fueron publicadas hasta la aparición de la informática, pero no estaban destinadas a la elaboración de un diccionario sino que fueron concebidas como base de una Gramática y, consecuentemente, el ordenamiento alfabético de las cédulas está subsumido al ordenamiento categorial.

En la primera mitad del siglo XX el distribucionalismo enfrenta la necesidad de estudiar lenguas completamente desconocidas, por lo tanto, el único elemento con el que cuenta el lingüista son cadenas de sonidos, lo que obliga a crear métodos de investigación completamente inductivos y empiristas, que prescinden de la tradición gramatical y ponen en el centro de la escena a los datos. La aplicación de la metodología distribucionalista a las lenguas ya conocidas lleva a postular que los datos son la única y suficiente realidad para el estudio del lenguaje.

Muchas descripciones del español sustentadas en el distribucionalismo se basan en material escrito; pero, dado que el objeto de estudio privilegiado son las manifestaciones orales, surge la necesidad de producir corpus utilizando grabaciones magnetofónicas, que eran los recursos los tecnológicos más avanzados en el momento del auge de esa corriente teórica. La mayoría de ellos se aplicó a investigaciones puntuales y forma parte de los papeles de trabajo no publicados del investigador; no obstante, en el año 1964, en el Segundo Simposio del Programa Interamericano de Lingüística y Enseñanza de Idiomas (PILEI), un grupo de investigadores presentó el "Proyecto de estudio del habla culta de las principales ciudades de Hispanoamérica", que consistía en grabar muestras de habla que luego fueron transcriptas y publicadas en distintos volúmenes. Aunque los datos recogidos provienen de la oralidad, este tipo de corpus los transforma en textos escritos y, en consecuencia, la recuperación de la información se debe hacer mediante una lectura lineal, lo que resulta bastante dificultoso.

[2] El trabajo de Cuervo iniciados como un proyecto individual en 1872 y fue completado en la década de 1990 por un equipo de investigadores del Instituto Caro y Cuervo de Bogotá.

[3] El *AGLE* de Fernández Ramírez fue digitalizado y está disponible en la página Web del Instituto Cervantes: http://www.cvc.cervantes.es/obref/agle/default.htm

Desde la perspectiva chomskiana lo que importa no son los enunciados en sí mismo sino las reglas generales y los procesos mentales que subyacen a ellos, de ahí que Chomsky (1965) ponga en tela de juicio el método inductivo del distribucionalismo con el argumento de que los datos no dan cuenta de la competencia del hablante sino de su actuación; además señala que, dado que los corpus tampoco pueden dar cuenta de la infinitud de los enunciados, resulta mucho más eficiente y económico recurrir a la introspección.

En las últimas décadas, el desarrollo de la Informática ha permitido la constitución de corpus lingüísticos de gran cobertura, dando lugar a la llamada Lingüística de corpus.

Los primeros proyectos de creación de corpus lingüísticos electrónicos surgieron el mundo anglosajón, pero actualmente existen varios corpus de lengua española disponibles en red[4]. Este tipo de corpus se pone a disposición de la lingüística una enorme cantidad de datos útiles para el estudio del lenguaje en distintas áreas.

Guillermo Rojo (2002) señala importantes diferencias teóricas y metodológicas entre los estudios empiristas del distribucionalismo y la Lingüística de corpus.

Desde una perspectiva teórica, la Lingüística actual no pretende hallar en el corpus todas las secuencias concretas ni los esquemas morfológicos y sintácticos posibles en una lengua; sin embargo, esto no implica que la introspección resulte una metodología más eficaz. En relación con la crítica chomskiana a los corpus lingüísticos, Rojo afirma:

> *Naturalmente, del acuerdo en que el objeto del trabajo lingüístico debe ser la totalidad de lo que es posible en una lengua (y no el conjunto de secuencias que la casualidad histórica ha incluido en un conjunto de mayor o menor tamaño) no se deduce forzosamente que la introspección sea el único procedimiento mediante el cual resulte posible acceder a los datos relevantes para la investigación y la comprensión de un determinado fenómeno lingüístico. La cuestión ha sido suficientemente debatida y hoy parece estar claro que, si bien los corpus no proporcionan todo lo que un lingüista necesita ni la caracterización estadística de los fenómenos es un elemento de consistencia universal por sí mismo, la pura introspección, aislada de los datos procedentes de los usos reales y debidamente documentados, está irremediablemente abocada a discusiones estériles sobre secuencias marginales o incluso imposibles, dejando sin explicación en muchos casos lo que realmente ocurre en una lengua o un determinado estado de una lengua. La competencia lingüística que tienen los hablantes es algo infinitamente más complejo de lo que se incorporaba a los primeros modelos chomskianos, de modo que el análisis de las secuencias producidas en condiciones reales resulta un elemento imprescindible del trabajo de los lingüistas. Es necesario, pues, como sucede en todas las ciencias, pasar continuamente de la teoría a los datos —a los datos reales, objetivos, no solo a los que el lingüista desdoblado en hablante pretende incorporar—, en aplicación estricta y bien entendida del método hipotético-deductivo.* (pág. 3)

Desde el punto de vista técnico los recursos informáticos permiten reunir mayor cantidad y variedad de datos y facilitan su recuperación de acuerdo con criterios mucho más rigurosos y exhaustivos, en tanto que se hace de forma automática.

En esta nueva perspectiva, los datos no son ni una ilustración del conocimiento gramatical, como ocurría en la Gramática Tradicional, ni la realidad absoluta del lenguaje, como pensaba el

[4] Sobre los diversos corpus electrónicos en inglés y en español ver Pérez Guerra (1999).

Distribucionalismo; sino un recurso que requiere estrictos controles. Por tal motivo, además de ser instrumento para el estudio del lenguaje, se constituyen en objeto de estudio de la lingüística, en tanto se debe garantizar que los textos incluidos sean cualitativamente representativos de la lengua que se pretende estudiar y estén adecuadamente codificados para que la información sea recuperable de manera automática.

El tema de los corpus como objeto de estudio de la Lingüística excede las perspectivas de mi trabajo; aquí me limito a describir brevemente las características del *Corpus de Referencia del Español Actual* (CREA)[5], uno de los bancos de datos lingüísticos elaborados por la Real Academia, *El Grial*[6], desarrollado por la Escuela Lingüística de Valparaíso y el *Corpus del Español*[7], creado por Mark Davies; e intento dar cuenta del empleo del CREA en la creación de una base de datos relacional específica para el estudio de las construcciones de *hacer* + infinitivo, y de los otros dos, como corpus de control, que permiten complementar y testear la relevancia los datos extraídos del CREA; y orientan acerca de cuál es el recurso más adecuado para el tratamiento informático del objeto de estudio.

## 2. Obtención de los datos

El CREA es un corpus de más de 150 millones de palabras distribuidas en documentos de todo el mundo hispanohablante, pertenecientes a distintos medios escritos y orales. Como señalé más arriba, fue creado por la RAE, por lo que puede considerarse el corpus "oficial" de la lengua española. Está organizado como una base de datos documental, disponible en forma gratuita por Internet para consultas lingüísticas, y codificado de modo tal que permite establecer criterios de búsqueda cronológico, geográfico, por autor, obra, tema y medio[8].

Dado que mi objeto de estudio son las construcciones de *hacer* + infinitivo; si pretendo elaborar un corpus específico, este debe incluir cualquier construcción que contenga estos elementos, solos o acompañados, y excluir cualquier otra construcción del verbo *hacer*, como podría ser, por ejemplo, *hacer* + sustantivo. El CREA, tal como está disponible actualmente en la red, no tiene codificación gramatical; por ello no es posible efectuar la búsqueda utilizando categorías lingüísticas: *hacer* e *infinitivo* solamente son accesibles como palabras, no como lema o etiqueta, respectivamente.

Para salvar esta dificultad, se ha recortado el objeto a construcciones en tercera persona del singular del pretérito perfecto simple del modo indicativo y a construcciones en primera persona del singular de igual tiempo y modo, que contengan proclíticos reflexivos; con este recorte la búsqueda se reduce a *hizo* más infinitivo y *me hice*.

La consulta con la entrada *me hice* arroja resultados imprecisos, porque incluye tanto construcciones donde a *me hice* le sigue un infinitivo como construcciones donde sigue cualquier otra categoría; pero, como el total de ejemplos recuperados no es demasiado numeroso, es posible seleccionar manualmente aquellos que son relevantes. En la consulta del 02/02/06 Se obtuvieron 456 casos de los cuales solo 19 resultaron relevantes (*me hice* + infinitivo).

La búsqueda con *hizo* es más dificultosa, ya que se obtienen 49738 casos, que, aunque se quisiera analizarlos manualmente, no sería posible hacerlo por limitaciones del CREA, que, cuando las consultas arrojan gran cantidad de resultados, no muestra los ejemplos.

[5] Cfr. www.rae.es
[6] Cfr. www.elgrial.cl/
[7] Cf.. http://www.corpusdelespanol.org.
[8] Para conocer detalles acerca de cómo está constituido y cómo se puede utilizar, ver el manual de consulta disponible en la misma página.

La alternativa es ir obteniendo estos casi cincuenta mil casos parcialmente, ya sea acotando los criterios (cronológico, geográfico, por autor, obra, tema y medio) o bien acotando la consulta. Con la finalidad de no perder ejemplos relevantes se efectuó la consulta: *hizo a** (con la cual se esperaba obtener, como resultado relevante, los ejemplos que contuvieran *hizo* más infinitivos comenzados en *a* y como resultado no relevante *hizo* con cualquier palabra comenzada en *a* que no fuese infinitivo), pero esta consulta fracasó ya que el resultado fue: *La consulta introducida es demasiado compleja, por favor simplifíquela.*

Finalmente, para la elaboración del proyecto de investigación, se decidió efectuar una consulta específica de *hizo* con 150 infinitivos comenzados en *a* (de *abalanzar* a *adquirir*), de los cuales solo 34 presentaron un total de 147 ejemplos y los demás no presentaron ninguno; el uso del asterisco al final del infinitivo permitió obtener también ejemplos con enclítico (ej. *abrir(se)*).

En un principio se trabajó con estos 166 ejemplos, pero posteriormente se realizaron nuevas búsquedas aplicando otras modalidades. Dado que es inevitable seleccionar manualmente los casos relevantes de los que no lo son, se acotó la búsqueda de *hizo* con criterio geográfico y se trabajó con los 4623 casos que se obtienen de documentos de la Argentina y los 672 que se obtienen de documentos de Uruguay. De estas búsquedas surgieron 915 nuevas construcciones de *hacer* + infinitivo.

Posteriormente se realizó una consulta teniendo en cuenta las posibilidades combinatorias de los clíticos, excluyendo los verbos relevados en las etapas anteriores, con este método se obtuvieron 73 nuevos casos; con estos ejemplos el corpus definitivo alcanza a 1154 ejemplos.

## 3. Organización y análisis gramatical de los datos obtenidos

Como se explicó, uno de los objetivo de mi proyecto de investigación es determinar las funciones sintáctico-semánticas que desempeñan los elementos contextuales de las construcciones de *hacer* más infinitivo. Por ejemplo, el clítico *lo* puede cumplir la función de sujeto o de objeto directo, lo que implica una ambigüedad funcional. En trabajos futuros se intentará establecer cuáles son las propiedades de los elementos léxicos que inciden en esta variaciones y con el análisis automáticos se hipotetizarán reglas combinatorias que permiten desambiguar, hasta cierto punto, las distintas construcciones.

La observación de los datos es un recurso insuficiente, pero indispensable, para lograr el objetivo perseguido. En este apartado se intenta explicar cómo se establece un método sistemático de organización de los ejemplos del CREA y del análisis gramatical que se realizó de ellos.

La organización de los datos es el punto de partida para establecer qué propiedades de lo constituyentes de las construcciones estudiadas resultan relevantes para su propia función o para la función de otro elemento. Por ejemplo, se puede determinar que algunas propiedades del infinitivo son determinantes para que el clítico “lo” sea únicamente sujeto en *Lo hizo caer*, o que pueda ser sujeto u objeto en *Lo hizo dormir*.

Dado el volumen de corpus obtenido, la creación de una base de datos relacional facilita enormemente la manipulación de los datos. Para satisfacer estas necesidades del proyecto se ha diseñado, en el programa Access de Microsoft.

Access es una base de datos relacional; este tipo de bases relacionan una **entidad** (objeto concreto o abstracto) o con un conjunto de **atributos**. A diferencia de las bases documentales, tienen restricciones en cuanto a las dimensiones de las entidades y el valor de sus atributos; se denomina **dominio** al conjunto de valores permitidos para cada atributo. Su utilidad consiste en que permite

organizar los ejemplos con distintos criterios, de modo que es posible agrupar aquellos donde uno o varios de sus constituyentes cumplen determinada función.

Además, el diseño de la base es un esquema para análisis manual de los ejemplos obtenidos:

Las entidades que conforman la base de datos son construcciones gramaticales (los ejemplos hallados); los atributos de esa entidad las categorías léxicas que conforman las construcciones (infinitivo; *hacer*; proclíticos *se, me, le, lo*; enclíticos *se, me, le, lo*; sintagmas nominales; sintagmas preposicionales con *a, de, por* y otras preposiciones[9]; adjetivo y adverbio).

Los dominios de la columna *Infinitivos* son el infinitivo que aparece en la entidad[10], los de la columna de *hacer*, la persona gramatical, y los de las demás columnas son las funciones sintácticas que se le asignan a cada categoría léxica (sujeto, objeto directo, objeto indirecto, índice de cuasi reflejo, complemento regido, predicativo y circunstanciales en posición argumental).

En el análisis sintáctico que se realiza de los ejemplos se aplican los siguientes criterios:

1) Se analizan únicamente los constituyentes que conforman el objeto de estudio.

2) La grilla tiene un orden fijo de elementos, que no siempre coincide con el que aparecen en el ejemplo.

3) Como consecuencia de lo anterior las oraciones interrogativas y relativas se analizan como si fuesen asertivas independientes.

4) Las oraciones de estilo directo y las subordinadas sustantivas se incluyen en la columna de los SN.

A continuación se muestran las columnas con los atributos y los primeros ejemplos:

Tabla 1: Base de datos relacional

| EJEMPLOS_TESIS | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **EJEMPLO** | **VERBO** | **HACER** | **SE** | **ME** | **LE** | **LO** | **E SE** | **E ME** | **E LE** | **E LO** | **SN** | **A** | **DE** | **POR** | **OT** | **ADJ** | **ADV** |
| La fuerte marejada le hizo abalanzarse hacia el timón | abalanzar | 3 | | | SUJ | | ICR | | | | | | | | CR | | |
| ...lo que les hizo abalanzarse sobre el ya prácticamente vacío baúl. | abalanzar | 3 | | | SUJ | | ICR | | | | | | | | CR | | |
| Su tosecilla me hizo abandonar la observación... | abandonar | 3 | | SUJ | | | | | | | OD | | | | | | |

Este diseño permite diversas consultas que simplifican el trabajo de asignación de propiedades. A continuación se muestran algunos ejemplos de estas consultas:

Si consultamos los casos en que el proclítico *lo* cumple la función de OD, obtenemos la siguiente tabla:

[9] En la tabla se agrupan los sintagmas preposicionales diferentes de *a*¸ *de, por*, porque no pueden cumplir función de sujeto.

[10] En las construcciones estudiadas, la función sintáctica de *hacer* y de los infinitivos no es ambigua, el primero será siempre núcleo de la oración subordinante y los segundos siempre serán núcleo de la oración subordinada, Se los incluye como entidades condicionan las posibles funciones de los otros elementos.

Tabla 2: Proclítico *lo* objeto directo

| **Consulta1** | |
|---|---|
| **EJEMPLO** | ***LO*** |
| Dispuesto a verla, la hizo abrir… | OD |
| ...siempre tomó la autoridad real para sí y la hizo acatar sin disputa por los otros. | OD |
| la hizo activar a control remoto por uno de sus cómplices | OD |

## 4. Análisis cuantitativo de los datos obtenidos

Después de haber extraído y analizados los casos del CREA del modo que se explicó más arriba, he tenido acceso a *El Grial* y a el *Corpus del Español* (*CdE*), que tienen la información etiquetada y lematizada, lo que simplifica enormemente la búsqueda.

*El Grial* es un corpus elaborado por el grupo ALADE, que dirige Givanni Parodi, en la Escuela Lingüística de Valaparaíso, dependiente de la Pontificia Universidad Católica de Valparaíso. Está disponible en línea y es de libre acceso[11]; según la información obtenida en el hipertexto *ver más* disponible en la presentación de "Búsqueda simple"[12], cuenta con más de 64 millones de palabras lematizadas y etiquetadas; en consecuencia, la búsqueda es mucho más eficaz: en pocos minutos se pueden obtener 12295 ejemplos de *hacer* con cualquier morfema flexivo y cualquier infinitivo u 841 ejemplos de *hizo* con cualquier infinitivo.

Las búsquedas en *El Grial* arrojan como resultado una planilla, que puede guardarse directamente como archivo de Excel, donde no aparece el contexto. Para obtener el contexto es necesario efectuar búsquedas parcializadas, que resultan un poco más dificultosas[13].

El *CdE* fue creado por Mark Davies en la Brigham Young University y cuenta con 100 millones de palabras provenientes de documentos del siglo XIII al siglo XX[14], aquí la información también está etiquetada y lematizada. Además, las búsquedas se pueden acotar por siglo; y los documentos correspondientes al siglo XX están subclasificados por tipos: *oral, ficcional, periodístico* y *académico* En pocos minutos se pueden obtener tanto las secuencias solicitadas como el ejemplo en su contexto. La única limitación es que no muestra más de mil resultados, por lo que también se hace necesario efectuar varias búsquedas parciales[15].

Rojo (2002) señala que en cualquier corpus un pequeño número de formas muy frecuentes constituye un alto porcentaje de la totalidad de las formas registradas. Esto implica que se da una

[11] Agradezco al Dr. René Venegas la paciencia y la buena predisposición con que me asesoró en el manejo de El Grial y la celeridad con la que respondió a todas mis consultas.
[12] Información obtenida el 29/06/2008. Remite a Giovanni Parodi "El Grial: interfaz computacional para anotación e interrogación de corpus en español"
[13] La última consulta fue realizada el 14/06/08.
[14] Hasta el momento (14/06/08) no he podido acceder a la información sobre los texto del corpus, que es uno de los capítulos del manual de ayuda en línea. Por lo tanto, no me resulta posible saber que porcentaje corresponde al siglo XX, que es el criterio con que se acotó la búsqueda.
[15] Se entiende por resultado el verbo *hacer* con un morfema flexivo y un infinitivo determinados; a continuación, se detalla cuántos ejemplos hay que contienen ese resultado. En el siglo XX, la parte más reducida que se puede consultar es por tipo de texto. En el caso de construcciones de *hacer* con cualquier morfema flexivo y cualquier infinitivo, todos los segmentos tienen menos de 1000 resultados, excepto el correspondiente a "ficción", que arroja 1000 resultados, lo que hace suponer que algunos quedaron excluidos.

tendencia a que pocas formas se repitan muchas veces y la ampliación del corpus aportará unas pocas formas nuevas y una gran reiteración de las ya existentes.

Con la finalidad de determinar con mayor exactitud cuál es la relación entre cantidad y variedad, se utilizan los datos obtenidos del *El Grial* y el *CdE* para proponer algunas comparaciones cualitativas con el corpus analizado.

### 4.1. Relación cantidad de ejemplos y cantidad de verbos

Según su propia definición, el CREA *responde a la intención de ofrecer a los investigadores de esta lengua y a los interesados en ella una muestra representativa y equilibrada del español estándar que se utiliza actualmente en el mundo*[16].

Pero en la constitución del corpus específico de construcciones de *hacer* + infinitivo a partir de esa base no se aplicó un criterio homogéneo: en algunos casos se consultó la totalidad del corpus, en otros se consultaron solo ejemplos en el subcorpus de la Argentina y Uruguay, que no se puede precisar con exactitud de cuántas palabras consta[17]. Además la búsqueda se limitó a construcciones con *hacer* en tercera persona y en primera persona con clíticos reflexivos.

*El Grial* permite obtener ejemplos donde *hacer* aparece con cualquier morfema flexivo, pero contiene datos provenientes, en un 90% de textos académicos obtenidos de los Programas de Estudio de cuatro carreras de la Pontificia Universidad Católica de Valparaíso (Psicología, Trabajo Social, Química Industrial e Ingeniería en Construcción)[18], lo que lo limita en lo que respecta a las variedades lingüísticas que registra.

A pesar de la imposibilidad de relacionar cuantitativamente los corpus específico con los corpus de origen; la comparación entre corpus de construcciones de *hacer* + infinitivo muestra tendencias que no parecen condicionadas por su origen. Con la finalidad de efectuar comparaciones cuantitativas con el corpus analizado (en adelante C1) se realizan las siguientes búsquedas:

a) En todo *El Grial*, **Lema:** *hacer*, **POS:** vbd infinitivo (en adelante C2). Se obtienen todos los casos donde aparece el verbo *hacer*, con cualquier morfema flexivo, seguido de un infinitivo.

b) En todo *El Grial*, **Forma:** *hizo*, **POS:** vbd infinitivo (en adelante C3). Se obtienen todos los casos donde aparece en verbo *hacer* en tercera persona del singular seguido de un infinitivo.

c) En la selección de los primeros documentos de *El Grial*, **Lema:** *hacer*, **POS:** vbd infinitivo (en adelante C4). Se seleccionó el corpus necesario para obtener al azar alrededor de mil ejemplos donde aparece el verbo *hacer*, con cualquier morfema flexivo, seguido de un infinitivo.

d) En el *CdE*, **Palabras:** *[hacer] [vr*]*, S20 (en adelante C5). Se obtienen los casos donde aparece *hacer* con cualquier morfema flexivo, seguido de infinitivo en los documentos correspondientes al Siglo XX. Se efectúa la búsqueda en cada tipo de texto para intentar que los resultados sean inferiores a 1000 (recuérdese que este es el tope que arroja el *CdE*), luego se los unifica.

[16] Cf.. http://corpus.rae.es/ayuda_c.htm#_Toc30228257 (04/06/08)

[17] El manual de ayuda señala que el 14% de las palabras corresponde al español rioplatense, que incluye la Argentina, Uruguay y Paraguay, lo que implica alrededor de 22 millones y medio de palabras; pero no ofrece datos cuantitativos precisos por país.

[18] Datos obtenidos de Giovanni Parodi "El Grial: interfaz computacional para anotación e interrogación de corpus en español", citado más arriba.

e) En el *CdE*, **Palabras:** *hizo [vr*]*, S20 (en adelante C6). Se obtienen todos los ejemplos donde aparece *hizo* seguido de infinitivo en documentos del Siglo XX.

### *4.1.1. Cantidad promedio de ejemplos por verbo*

Se establece una comparación entre la cantidad promedio de ejemplos entre C1 y los datos que aportan las búsquedas anteriores y se obtienen los siguientes guarismos:

Tabla 3: Cantidad de verbos por ejemplo

| **Corpus** | **Cant de ejs.** | **Cant de verbos** | Ejs por verbos (prom) |
|---|---|---|---|
| C1 | 1154 | 334 | **3,45** |
| C2 | 12295 | 841 | **14,62** |
| C3 | 935 | 238 | **3,93** |
| C4 | 1018 | 268 | **3,80** |
| C5 | 5015 | 673 | **7,45** |
| **C6** | **1119** | **371** | **3,01** |

### *4.1.2. Concentración de ejemplos por verbo*

Si se observa la concentración de ejemplos, se obtienen los siguientes resultados:

Tabla4: Concentración de ejemplos

| **Corpus** | **% ejemplos** | Concentración en % de verbos |
|---|---|---|
| C1 | 90,02% | **65,56%** |
| C2 | 90,02% | **25,92%** |
| C3 | 90,06% | **60,92%** |
| C4 | 90,08% | **62,31%** |
| C5 | 90,01% | **37,89%** |
| **C6** | **90,00%** | **69,81%** |

## 4.2. Relación entre variaciones morfológicas de hacer y cantidad de patrones sintácticos

En este apartado se analiza la incidencia que tiene la variación morfológica de *hacer* en los patrones sintácticos que muestran los verbos. Adicionalmente, se observa si la aparición de nuevos patrones está condicionada por la morfología de *hacer* o por la ampliación del corpus.

Se asume que dos ejemplos muestran el mismo patrón sintáctico cuando tienen los mismos constituyentes en las mismas funciones, independientemente del orden en el que aparezcan. Por el contrario tienen distintos patrones sintácticos cuando en uno de ellos aparece algún elemento que no aparece en el otro o cuando aparece el mismo elemento con distinta función.

Para esta determinación se aplicó el siguiente procedimiento:

1) Se creó un corpus de control (CC) mediante consultas en el *CdE* con *hacer* con cualquier morfema flexivo finito de cuatro verbos que en C1 presentan una cantidad importante de ejemplos y tienen diferentes propiedades:

a) *saber*: es transitivo, no selecciona OI, no tiene construcciones cuasi reflejas, no rige complementos con preposición.

b) *llegar*: selecciona OI, no tiene construcciones cuasi reflejas, puede regir complementos con preposición.

c) *caer*: es intransitivo, no selecciona OI, puede tener construcciones cuasi reflejas, puede regir complementos preposicionales.

d) *olvidar*: es transitivo o intransitivo, no selecciona OI. Si es transitivo, no rige complemento con preposición. Si es intransitivo, tiene construcciones cuasi reflejas o rige complementos con preposición.

2) Se realizaron consultas en Access de cada uno de estos verbos en la base de datos C1 y CC.

Las consultas incluyen todos los campos de la tabla, salvo EJEMPLO y H, se acota el campo infinitivo con el verbo que se quiere mostrar, los demás se ordenan en forma ascendente y en "propiedades" se marca la opción "valores únicos". Esta opción suprime las repeticiones de los registros que tienen valores idénticos, de este modo se obtiene lo que definí como patrones sintácticos de cada verbo independientemente del ejemplo concreto y de la persona de *hacer*[19]:

| Verbo | C1 | | CC | | Coinc | %Coinc |
|---|---|---|---|---|---|---|
| | Ejs | Patrones | Ejs | Patrones | | |
| ***saber*** | 77 | 15 | 64 | 12 | 10 | 58,82% |
| ***llegar*** | 49 | 18 | 68 | 18 | 10 | 38,46% |
| ***caer*** | 14 | 9 | 33 | 13 | 7 | 53,33% |
| ***olvidar*** | 12 | 6 | 32 | 8 | 4 | 40,00% |

## 5. Conclusiones

C1, C3, C4, C6 resultan muy similares en cuanto a la cantidad de datos y a la proporción que existe entre ellos.

En C2 una multiplicación cuantitativa que excede el décuplo de ejemplos apenas triplica la cantidad de verbos de C3 y C4 y no alcanza a triplicar la de C1 y C6.

La multiplicación de ejemplos próxima al quíntuplo en C5 se aproxima a la duplicación de verbos de C1, C3, C4 y C6.

Estos datos muestran una acumulación de ejemplos que contienen el mismo verbo y una gran concentración de ejemplos en unos pocos verbos.

[19] Se excluye el campo *Ejemplo* porque éste es un valor único en todos los casos, pero no forma parte del patrón sintáctico.

Además, permiten concluir que la variedad lingüística del corpus de origen no es determinante para la relación entre cantidad de ejemplos y cantidad de verbos porque, de ser así, C2 y C3, por un lado y C5 y C6, por el otro, tendrían que mostrar la misma tendencia; hecho que no se verifica en la comparación.

Por su parte, las diferencias entre C2 y C4, muestran que las propiedades morfológicas de *hacer* tampoco tienen mayor relevancia en esta relación cuantitativa.

De modo que la única variable que parece tener incidencia en la proporción de los datos es el volumen total del corpus. De estas relaciones, se puede inferir que para duplicar la cantidad de verbos habría que multiplicar por cinco la cantidad de ejemplos y para triplicarla sería necesario multiplicarlos, al menos, por diez; lo que excedería las posibilidades de un trabajo individual.

La ampliación de ejemplos analizado no revela nuevas propiedades en la rección de los verbos[20], ni nuevas propiedades funcionales en los clíticos y los sintagmas analizados, ni tampoco incidencia del morfema flexivo de *hacer* en la construcción de infinitivo. El escaso porcentaje de coincidencias se debe a diferentes combinaciones de las mismas funciones sintácticas.

La regularidad de las propiedades rectoras de los verbos y de la funciones de los demás constituyentes relacionados con ellos hace suponer que un autómata de estados finitos, como XFST, resulta adecuado para el tratamiento informático porque es muy probable que las construcciones no incluidas en C1, aunque pudieran presentar distintas combinaciones de complementos, tuvieran las mismas propiedades básicas que estas. Para analizar automáticamente construcciones con verbos que no se encuentran en C1 lo único que habría que hacer es caracterizar adecuadamente el verbo.

En cambio, la variedad de las posibilidades combinatorias hace dificultoso el empleo de un programa como MPS, donde es necesario declarar una a una las combinaciones posibles.

## Referencias

[1] Cano Aguilar, Rafael (1977): “Las construcciones causativas en español” en *Boletín de la Real Academia Española.* Tomo LVII, Cuadernos CCXI, mayo – agosto de 1977 (págs. 221 – 258) y CCXII, septiembre – diciembre de 1977 (págs. 323 – 351).

[2] Chomsky, Noam (1965). *Aspectos de la teoría de la sintaxis,* Madrid, Aguilar.

[3] Couto, J., G. Crispino, M. Grassi, M. Skorodynski (1999): "Estructuración de índices gramaticales y léxicos para la extracción y recuperación de Información". XV Congreso de la [3] Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). Lérida, España, 1999. http://www.fing.edu.uy/~jcouto/

[4] Cuervo, Rufino José (1994): *Diccionario de construcción y régimen de la lengua castellana*; continuado y editado por el Instituto Caro y Cuervo, Santafé de Bogotá, Instituto Caro y Cuervo.

[5] Davies, Mark (2002): Corpus del español (100 millones de palabras, siglo XIII - siglo XX). Disponible en http://www.corpusdelespanol.org.

[6] Escuela Lingüística de Valparaíso: *El Grial.* http://www.elgrial.cl

[20] En búsquedas no sistemáticas se hallaron ejemplos como *los hizo saberse iguales* e *hizo saber de su disconformidad*, que muestran otras posibilidades en la construcción de *saber*. Esto es predecible porque, como se dijo más arriba, ningún corpus, por extenso que sea puede cubrir todas las posibilidades de una lengua.

[7] Real Academia Española: Banco de datos: (CORDE) [en línea] *Corpus diacrónico del español.* (http://www.rae.es) y (CREA) [en línea] *Corpus de referencia del español actual.* (http://www.rae.es)

[8] Fernández Ramírez, Salvador (1997-2008): *Archivo Gramatical de la Lengua Española.* http://www.cvc.cervantes.es/obref/agle/default.htm

[9] Grassi, M., M. Malcuori, J. Couto, J. Prada, D. Wonsever (2001): "Corpus informatizado: textos del español del Uruguay (CORIN)". SLPLT-2 - Second International Workshop on Spanish Language Processing and Language Technologies. Jaén, España, septiembre 2001. http://www.fing.edu.uy/~jcouto/

[10] Pérez Guerra, Javier (1999) “Estándares de anotación en lingüística de corpus” en Revista española de lingüística aplicada, ISSN 0213-2028, Vol. 1, 1999 (Ejemplar dedicado a: Panorama de la investigación en Lingüística Aplicada), Págs. 25-52. http://dialnet.unirioja.es/servlet/articulo?codigo=227024

[11] Rodríguez Yunta, Luis (2001): Bases de datos documentales: estructura y uso, en: Maldonado, Ángeles (coord.). *La información especializada en Internet*. Madrid: CINDOC, 2001

[12] Rojo, Guillermo (2002): “Sobre la Lingüística basada en el análisis de corpus” http://www.uzei.com/Modulos/UsuariosFtp/Conexion/archivos54A.pdf