Corpus y construcciones

Perspectivas hispánicas

Edición a cargo de
Marta Blanco
Hella Olbertz
Victoria Vázquez Rozas

Verba
Anexo 79

2019

Servizo de Publicacións e Intercambio Científico • UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CORPUS Y CONSTRUCCIONES
PERSPECTIVAS HISPÁNICAS

Edición a cargo de
Marta Blanco
Hella Olbertz
Victoria Vázquez Rozas

Verba
Anexo 79

2019
Universidade de Santiago de Compostela

Maquetación
Antón García

Edita
Servizo de Publicacións
e Intercambio Científico
Campus Vida
15782 Santiago de Compostela
www.usc.es/publicacions

DOI https://dx.doi.org/10.15304/9788417595876
ISSN 2341-1198
ISBN 978-84-17595-87-6

ÍNDICE

PRESENTACIÓN

Los trabajos reunidos en *Corpus y construcciones. Perspectivas hispánicas* derivan de las contribuciones de sus autores al encuentro científico del mismo nombre celebrado en la Facultade de Filoloxía de la Universidade de Santiago de Compostela los días 22 y 23 de noviembre de 2018. El evento, organizado por el grupo de investigación Gramática del español, sirvió como marco para presentar un conjunto de investigaciones novedosas relacionadas con la lingüística de corpus que impulsaron un estimulante debate entre los participantes. Prueba del interés del encuentro y de la calidad de los trabajos expuestos es que una selección de diez aportaciones, en versión escrita ampliada y enriquecida, se compartan ahora de forma permanente con la comunidad académica a través de la edición de este anexo de la revista *Verba*.

El volumen integra trabajos relacionados con tres líneas preferentes de investigación del grupo Gramática del español desde su constitución: el análisis tanto sincrónico como diacrónico de estructuras gramaticales, las relaciones entre gramática y léxico, y la elaboración de corpus y bases de datos lingüísticas. Una fructífera combinación de estas tres orientaciones dio lugar en los años noventa del siglo XX a la construcción de la Base de datos sintácticos del español www.bds.usc.es, un recurso electrónico pionero elaborado a partir del análisis de un corpus de textos orales y escritos de las dos décadas anteriores (corpus ARTHUS: www.bds.usc.es/corpus.html), enriquecido posteriormente con información semántica y léxica en el proyecto ADESSE en la Universidad de Vigo http://adesse.uvigo.es/.

Desde el núcleo inicial de estudios centrados en las estructuras sintácticas, las investigaciones de los miembros del grupo se han ampliado con el análisis de nuevos problemas y la adopción de nuevos enfoques. Cabe destacar el creciente interés por los fenómenos de contacto, por un lado, y por el contraste entre lenguas y variedades, por otro. En ambas perspectivas se atribuye una relevancia crucial a los datos de uso como fundamento empírico de la investigación.

La importancia concedida al uso lingüístico real, oral y escrito, como objeto imprescindible de una aproximación funcional al estudio de la lengua no solo ha estimulado las investigaciones basadas en materiales auténticos, sino que ha evidenciado la necesidad de disponer de corpus que proporcionen la información requerida en las condiciones de accesibilidad adecuadas. En los últimos años, el trabajo del grupo sobre corpus se ha ampliado con la creación de nuevos recursos y nuevas herramientas de tratamiento y anotación de textos que facilitan su explotación tanto en lingüística teórica y descriptiva como en el campo de la lingüística aplicada.

CONTENIDO DEL VOLUMEN

El volumen tiene dos partes, que se relacionan respectivamente con los dos ámbitos de desarrollo de la lingüística de corpus; por un lado, con el análisis de fenómenos lingüísticos basados en datos extraídos de corpus y, por otro, con el diseño, elaboración y enriquecimiento de corpus con vistas a una adecuada recuperación y explotación de la información que contienen.

En cuanto a las lenguas sobre las que versan los trabajos, aunque el español es objeto de estudio de buena parte de ellos, el volumen incluye aportaciones sobre el gallego (capítulos 6 y 7), el inglés (capítulo 1), el portugués (capítulo 5) y el alemán (capítulo 10). Además, varios capítulos muestran la necesidad de un enfoque plurilingüe, bien para dar cuenta de fenómenos de variación y cambio en situaciones de contacto, como ocurre con el español y el inglés en Nuevo México (capítulo 1), bien para desarrollar recursos lingüísticos para la enseñanza de lenguas extranjeras o para la traducción, como el corpus de aprendices CAES (capítulo 9) o el corpus paralelo alemán/español PaGeS (capítulo 10).

La primera parte, dedicada a los estudios gramaticales con datos de corpus, se abre con el capítulo de Rena Torres Cacoullos y Catherine Travis «Gramáticas en contacto en un corpus bilingüe», cuyo objetivo es revisar la validez de la hipótesis de la convergencia gramatical, es decir, determinar si el cambio lingüístico se debe realmente al contacto y si el estatus minoritario o dominante de cada lengua determina el sentido del cambio. El estudio se fundamenta en la observación sistemática de las muestras de habla espontánea que integran el *Corpus bilingüe español-inglés de Nuevo México*, que por su configuración constituye un recurso idóneo para comparar el uso que hacen de ambas lenguas los hablantes bilingües de la comunidad objeto de estudio. Mediante una metodología cuantitativa cuidadosamente diseñada, que confronta los datos del corpus bilingüe con la información obtenida de dos corpus monolingües de español e inglés, se establece el grado de semejanza

entre las gramáticas bilingües de inglés y español y entre las correspondientes gramáticas monolingües en tres tipos de estructuras variables: las perífrasis progresivas, el uso de indicativo y subjuntivo en cláusulas subordinadas sustantivas y la expresión variable del sujeto pronominal. Los resultados del estudio contradicen las propuestas anteriores de convergencia gramatical y sustentan la continuidad de la independencia de las dos gramáticas de los hablantes bilingües.

El segundo capítulo, elaborado por Anton Granvik y titulado «Sobre los orígenes de la construcción encapsuladora en español», se basa en los datos del corpus del *Nuevo diccionario histórico del español* para ofrecer un detallado análisis diacrónico de la construcción por la que ciertos sustantivos abstractos como *causa*, *condición* o *idea* remiten a una unidad de información compleja de tipo proposicional. Para superar las dificultades que plantea el criterio de la identidad experiencial en la determinación de la función encapsuladora de cada sustantivo, el estudio parte de una interpretación formal y esquemática de la construcción, compatible con el enfoque construccionista, e incorpora tres propiedades gramaticales como criterios operacionales de encapsulación: la determinación de la frase nominal, su función sintáctica y el tipo de unidad introductora. Esta aproximación metodológica permite al autor establecer una escala de tipicidad a partir de una amplia muestra de usos de nueve sustantivos entre los siglos XIII y XX. El estudio se completa con un minucioso análisis cualitativo semántico-cognitivo y textual de los elementos seleccionados que permite detectar diferencias funcionales condicionadas por la semántica léxica propia de cada sustantivo.

La gramática de construcciones y la lingüística de corpus constituyen el marco teórico y metodológico del capítulo 3, «*Entre miradas de asombro*: aportaciones de la Lingüística de Corpus al estudio de una construcción con la preposición *entre*», de Belén López Meirama y Carmen Mellado Blanco. En este caso la base empírica del análisis se extrae del CORPES XXI a partir de una búsqueda inicial de proximidad a la que se aplica un filtrado manual para seleccionar la combinación [*entre* + sustantivo$_{\text{plural/corporal}}$]. El estudio detallado de todas las secuencias que presentan tal estructura abarca tanto los aspectos morfológicos y sintácticos como sus valores semánticos y pragmáticos, con especial atención a la unidad léxica variable de la construcción, identificada como un sustantivo, plural o en coordinación, de comunicación o expresión corporal. El análisis revela el predominio de una configuración prototípica de la unidad fraseológica en torno a un número reducido de sustantivos nucleares que dan cuenta del 50% de los casos. Se observa asimismo la existencia de un efecto de coerción semántica ejercido por el primer sustantivo coordinado

sobre el segundo. Como resultado de alcance más general, el trabajo ofrece el diseño y la aplicación de una propuesta metodológica extensible a otras unidades fraseológicas preposicionales.

En el capítulo 4, «En torno al concepto de *perfil combinatorio*», Inmaculada Mas Álvarez realiza un recorrido por diferentes propuestas que han ido configurando la noción de 'perfil combinatorio' como una aportación clave para el enriquecimiento de las descripciones lexicográficas a partir de los resultados de la lingüística de corpus. El concepto se fundamenta en que los datos de frecuencia y coocurrencia léxico-gramatical obtenidos a través de concordancias permiten identificar, con base en el uso real, los patrones constructivos de las unidades analizadas. En el texto se describen las características de algunos recursos que incorporan de manera sistemática información sobre la combinatoria sintáctica verbal, como la BDS y ADESSE, este último incluyendo propiedades semánticas y definiciones léxicas. Se informa asimismo sobre las opciones que ofrece la herramienta *Sketch Engine* para analizar en detalle el perfil combinatorio de los elementos léxicos y establecer comparaciones entre perfiles; y finalmente se resume el método del *collostructional analysis*, que mide el grado de atracción entre unidades léxicas y construcciones y ha demostrado su utilidad en el análisis de diversas relaciones léxicas en diferentes lenguas.

Cierra la primera parte del volumen el capítulo de Hella Olbertz «Funciones pragmáticas en el portugués brasileño: un enfoque discursivo-funcional». La autora parte de la comparación entre el español y el portugués para examinar con detalle la expresión de las funciones pragmáticas de tópico y foco en el portugués brasileño. El estudio se basa en un detallado análisis cualitativo del uso registrado en corpus orales comparables: PRESEEA de Alcalá de Henares para el español, *Iboruna* para el portugués del Brasil y C-ORAL-ROM y una parte de *Português Falado* para la variedad de Portugal. Los conceptos funcionales empleados en el análisis —tópico y foco, agente y paciente, sujeto y objeto— se definen y contextualizan en el modelo de la gramática discursivo-funcional, del que se ofrece una breve pero ilustrativa presentación. En el núcleo del trabajo se explican de forma pormenorizada y empíricamente fundamentada (i) los cambios que ha experimentado el portugués brasileño en la expresión personal del sujeto y (ii) cómo la progresiva sobrecarga funcional de la concordancia verbal de 3ª persona de singular provocó la generalización del pronombre sujeto, cuya desemantización y pragmaticalización ha dado origen a una marca gramatical de la función de tópico. Se establece así un contraste entre el español, lengua con una marca propia de la función focal, y el portugués de Brasil, que ha desarrollado un mecanismo innovador para identificar el tópico.

La segunda parte del volumen integra cinco capítulos centrados en el diseño y desarrollo de corpus. En el primero de ellos, el capítulo 6, Eva María Domínguez Noya, María Sol López Martínez y Francisco Mario Barcala Rodríguez presentan «Corpus de Referencia do Galego Actual (CORGA): composición, codificación, etiquetaxe e explotación». El trabajo informa en primer lugar de la composición y estructuración interna del corpus, que abarca desde 1975 hasta la actualidad y alcanza en la versión 3.1 una extensión de 40 178 271 palabras. Los materiales del CORGA consisten en una amplia variedad de textos escritos y en una muestra oral de 25 horas de emisiones radiofónicas transcritas y alineadas con el audio. Se destaca la importancia del diseño del corpus, tanto en lo que atañe a los criterios de clasificación textual (fecha, tipo o género, área temática) como en lo referente al tratamiento de la variación gráfica —que se resuelve introduciendo la categoría de 'hiperlema'— y morfológica. En el capítulo se expone el protocolo que siguen los documentos en el proceso de construcción del corpus y se presenta el sistema de etiquetación morfosintáctica llevado a cabo con el etiquetador del gallego actual XIADA, desarrollado en relación con el corpus CORGA y adaptado a sus necesidades. El trabajo se completa con la descripción del sistema de recuperación de los datos del corpus a través de la aplicación de consulta.

A continuación, en el capítulo titulado «CORILGA: un corpus para estudar a variación e o cambio do galego falado», Elisa Fernández Rei y Xosé Luís Regueira contextualizan la creación de un corpus oral actual en el marco de las iniciativas impulsadas por el *Instituto da Lingua Galega* (ILG) a lo largo de las últimas décadas. Buena parte de las 105 horas de grabación que recoge el corpus proceden de proyectos anteriores y forman un valioso conjunto de materiales de habla que se ponen ahora a disposición pública en las condiciones de acceso y consulta adecuadas para facilitar el estudio de la variación diastrática, diafásica e incluso diacrónica, ya que los registros sonoros se extienden desde 1965 hasta el momento actual. Para la transcripción y anotación del corpus en diferentes niveles se emplea el programa ELAN, que permite integrar diferentes recursos de tecnología del habla desarrollados en colaboración con el Grupo de Tecnoloxías Multimedia de la Universidade de Vigo, entre las que destacan las herramientas de alineación texto-voz, reconocimiento de voz y transcripción automática.

El capítulo 8, elaborado por Eva M.ª Domínguez Noya, Raquel Rivas Cabanelas, M.ª Paula Santalla del Río y Rebeca Villapol Baltar, trata de «Problemas afrontados en la etiquetación morfosintáctica del corpus ESLORA». Tras resumir las condiciones de etiquetación de varios corpus orales desarrollados con anterioridad, el trabajo ofrece información sobre los recursos

utilizados en el tratamiento de ESLORA, un corpus que recoge entrevistas y conversaciones de hablantes de español en Galicia. Se presentan las características del etiquetador, el etiquetario, el diccionario y el corpus de entrenamiento, y se describen las fases del proceso de etiquetación. El núcleo del capítulo aborda algunas de las dificultades que se plantearon en el proceso de revisión manual de los resultados de la etiquetación automática junto con las soluciones adoptadas en cada caso. Se explican, entre otras, las opciones elegidas para la anotación de los numerales, los ruidos comunicativos, los conectores discursivos, el tratamiento de formas que se apartan del estándar normativo y ciertos usos de formas frecuentes como *que* y *tal*. Como conclusión, el trabajo subraya la necesidad de adaptar los recursos y las decisiones de anotación a las particularidades de los materiales anotados, en este caso a las características específicas de un corpus oral como ESLORA.

«El *Corpus de Aprendices de Español* (CAES) y sus aplicaciones para la enseñanza/aprendizaje del español como lengua extranjera» es el título del capítulo 9, en el que Ignacio Palacios Martínez, Francisco Mario Barcala Rodríguez y Guillermo Rojo describen las características del corpus, las fases de su construcción y etiquetación e ilustran con casos prácticos las principales líneas de explotación de la información que contiene. En su versión 1.2 de agosto de 2018 el CAES comprende cerca de 600 000 elementos lingüísticos correspondientes a la producción escrita de una amplia muestra de aprendientes de español con niveles desde A1 a C1 hablantes iniciales de alguna de las siguientes lenguas: árabe, chino mandarín, francés, inglés, portugués y ruso. El hecho de que la anotación morfosintáctica de los textos haya sido rigurosamente desambiguada de forma manual incrementa notablemente la utilidad del recurso, tanto como fuente de información para incidir de forma efectiva en el proceso de aprendizaje, como por las posibilidades de aprovechamiento de los datos en la elaboración de actividades de aula y materiales didácticos. El trabajo muestra asimismo que el corpus, por sus características y por las facilidades de consulta que ofrece, constituye un recurso valioso en la formación del profesorado de ELE y como fuente de referencia para el diseño curricular.

En el capítulo final del volumen, que lleva por título «Multifuncionalidad de los corpus paralelos, ejemplificada con el corpus alemán / español PaGeS», Irene Doval y Tomás Jiménez dan cuenta de la composición del corpus PaGeS, del proceso de compilación y de las diferentes opciones de recuperación de datos que ofrece. PaGeS es un corpus formado en su parte nuclear por textos fundamentalmente narrativos alemanes y españoles escritos en las últimas décadas y alineados con sus traducciones al español y al

alemán. La versión 2.0 de abril de 2019 contiene 28 millones de unidades distribuidas de forma equilibrada entre ambas lenguas. En el capítulo se describen los pormenores de la preparación de los materiales y las dificultades de segmentación y alineado de los textos junto con las soluciones alcanzadas, lo que implica, por una parte, la selección de herramientas computacionales adecuadas, y por otra, un laborioso trabajo de validación manual. La presentación se completa con la ilustración de las posibilidades que ofrece el motor de búsqueda para recuperar la información contenida en el corpus. Se apuntan finalmente las previsiones de enriquecimiento y mejora del recurso en un futuro inmediato, como la ampliación del alineado al nivel de la palabra y la etiquetación morfosintáctica de los textos.

AGRADECIMIENTOS

Tanto la edición del presente volumen como la organización del encuentro en el que se presentaron versiones previas de los trabajos aquí incluidos se enmarcan en la estrategia de consolidación propuesta por el grupo de investigación Gramática del español para el trienio 2017-2019. En este período el grupo fue beneficiario de una ayuda del Programa de consolidación y estructuración de unidades de investigación competitivas, en la modalidad de Grupos con potencial de crecimiento, concedida por la Consellería de Cultura, Educación e Ordenación Universitaria y la Consellería de Economía, Emprego e Industria de la Xunta de Galicia a través de la Axencia Galega de Innovación (ref. nº ED431B 2017/39). Para la organización del encuentro contamos asimismo con la financiación del proyecto de investigación ESLORA+ (ref. FFI2017-86379-P) subvencionado por la Agencia Estatal de Investigación (AEI) y por el Fondo Europeo de Desarrollo Regional (FEDER).

Por último, deseamos agradecer públicamente la colaboración generosa de las personas que elaboraron los dictámenes de los trabajos enviados para publicación en el volumen. Su contribución ha sido imprescindible para garantizar la calidad de los diez capítulos finalmente seleccionados, cuyas versiones definitivas se han beneficiado de los pertinentes comentarios aportados en los informes de evaluación.

Marta Blanco, Hella Olbertz, Victoria Vázquez Rozas

GRAMÁTICAS EN CONTACTO EN UN CORPUS BILINGÜE

Grammars in contact in a bilingual corpus

Rena Torres Cacoullos
Universidad Estatal de Pensilvania
Catherine E. Travis
Universidad Nacional de Australia

Resumen

Es bastante común la idea de que el contacto de lenguas conduce a la convergencia gramatical, a pesar de la escasa evidencia. Para comprobar si se ha producido un cambio y que este se debe al contacto, hay que cumplir con dos requisitos: uno social, el de mostrar que el cambio es un patrón regular en una comunidad de habla, y otro lingüístico, el de distinguir la influencia externa de las tendencias lingüísticas internas. Para cumplir con este par de requisitos, se utiliza un *corpus bilingüe* que permite contextualizar las estructuras de interés tanto con respecto a la comunidad de habla como con respecto al sistema gramatical. En este artículo se describe la construcción de un corpus bilingüe en el que hay suficiente representación de las dos lenguas en contacto para aplicar un *criterio de convergencia* basado en los patrones de variación interna entre formas morfosintácticas que aparecen en alternancia. Las comparaciones abarcan tanto las variedades bilingües como variedades monolingües como punto de referencia. Se ejemplifica con el español en contacto con el inglés, en una comunidad bilingüe establecida.

Palabras clave: contacto lingüístico, convergencia gramatical, continuidad lingüística, comunidad bilingüe, corpus bilingüe

Abstract

It is widely held that language contact leads to grammatical convergence. The evidence for this, however, remains scant. To verify that language change has occurred and that

it is due to contact, two requirements must be met, one social —to demonstrate that the change has become a regular pattern in a speech community— and one linguistic —to distinguish external influence from internal linguistic tendencies. These twin requirements are met by a *bilingual corpus* that permits contextualization of linguistic forms with respect to both the speech community and the grammatical system. This article describes the construction of a bilingual corpus with each contact language sufficiently represented to apply a *convergence criterion* based on patterns of internal variation between alternative morphosyntactic forms, as compared across bilingual varieties and monolingual benchmarks. We illustrate with Spanish in contact with English in an established bilingual community.

Keywords: language contact, grammatical convergence, linguistic continuity, bilingual community, bilingual corpus

1. Introducción

Una hipótesis de difusión amplia es que el bilingüismo conlleva cambio en por lo menos una de las lenguas en contacto, normalmente en la lengua minoritaria[1]. La idea es que el contacto resulta en una especie de mezcla de lenguas, tal como lo implican términos como *Spanglish* en los Estados Unidos, *Franglais* en Canadá o *Türkendeutsch* en Alemania. En cuanto a la morfosintaxis, la hipótesis de la *convergencia gramatical* sostiene que dos gramáticas en contacto no se mantienen independientes, sino que se van convirtiendo en una. Así, por ejemplo, en su muy citado libro, Thomason y Kaufman proponen que los bilingües, al alternar entre lenguas, favorecen estructuras compartidas por sus dos lenguas, lo cual conduce a lo que llaman «ajuste interlingüístico» («cross-language compromise») (Thomason y Kaufman 1988: 154). Dicho de otra forma, los bilingües «integran principios gramaticales de cada sistema lingüístico» («are integrating grammatical principles from each linguistic system») (Goldrick *et al.* 2016: 860). Pero si bien se documenta el cambio léxico, e inclusive fonético, la evidencia para el cambio morfosintáctico es más bien escasa (*cfr.* Poplack & Levey 2010).

¿Cómo comprobar una hipótesis de cambio lingüístico atribuido al contacto? La pregunta tiene dos componentes. En primer lugar, ¿cómo saber si se trata realmente de cambio, y no de idiosincrasia o desplazamiento lingüístico al nivel de individuos? Y, una vez establecido el hecho del cambio, ¿cómo saber si este se debe al contacto, y no a un mecanismo interno? La respuesta,

[1] Se agradece el apoyo del National Science Foundation (NSF), BCS 1019112/1019122 y BCS 1624966 y del Australian Research Council (ARC) Centre of Excellence for the Dynamics of Language. Le agradecemos a Manuel Delicado Cantero la generosa lectura y a los evaluadores sus observaciones.

por su parte, tiene dos requisitos: uno social y otro lingüístico. El requisito social es que se debe mostrar que un cambio sea un patrón regular en una comunidad de habla porque, aunque una innovación lingüística puede ser producida por cualquier hablante, no la consideramos un cambio hasta que no sea transmitida a otros hablantes. El requisito lingüístico es que hay que hacer una distinción entre el cambio interno, que sigue las tendencias interlingüísticas, y el cambio externo, que sigue alguna particularidad gramatical de otra variedad lingüística con la que se entra en contacto (*cfr.* Blas Arroyo 1999: 417).

Cumplimos con este par de requisitos mediante la construcción de un *corpus bilingüe* que capta ambas lenguas habladas en la comunidad bilingüe y la aplicación de un *criterio de convergencia* basado en la comparación de los patrones de estructuras paralelas en las dos lenguas.

En este artículo describimos los datos y la metodología para investigar las consecuencias lingüísticas del contacto de lenguas. Los resultados obtenidos muestran que el cambio no es una consecuencia inevitable del contacto y sugieren que, al contrario, el uso regular y alternante de dos lenguas puede conllevar continuidad.

2. El criterio de convergencia: comparar la variación lingüística interna en lenguas minoritarias y mayoritarias

Un debate fundamental en torno a las situaciones de contacto lingüístico es si conllevan necesariamente cambio. En particular, se pregunta si el contacto da lugar a la convergencia gramatical, es decir, si los bilingües, en vez de conservar dos gramáticas independientes, las mezclan en una sola gramática convergente.

La mayoría de las propuestas de convergencia gramatical se basa en una comparación entre la variedad bilingüe y una variedad referente de la lengua minoritaria, que puede ser una monolingüe, por ejemplo, del país de origen (como entre el turco en los Países Bajos y en Turquía), o una menos bilingüe, por ejemplo, de contacto menos intenso con la lengua mayoritaria o de una etapa anterior al contacto (por ejemplo, entre el gallego urbano y el rural en Galicia; entre el vasco de hogares en que es la lengua mayoritaria y de aquellos en que no lo es; o entre el español de la segunda y de la primera generación de inmigrantes en los EEUU) (*cfr.* Silva-Corvalán 1994; Dubert García 2005; Doğruöz & Backus 2007; Rodríguez Ordóñez 2017). Se podría decir que se trata de una comparación vertical (figura 1, flecha vertical a la izquierda). Por otra parte, ha sido mucho menos común que se realice la

comparación correspondiente para la lengua mayoritaria (una excepción es Poplack *et al.* 2012) (por ejemplo, el inglés de los hispanohablantes con el inglés de los angloamericanos; flecha vertical a la derecha).

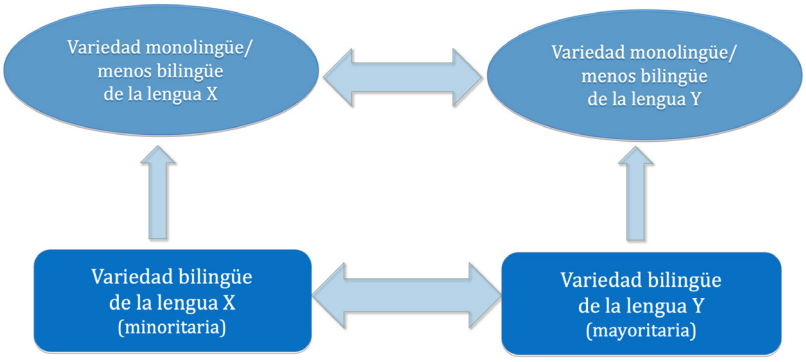


FIGURA 1. Criterio de convergencia: comparaciones de corpus

El criterio de convergencia que proponemos privilegia las comparaciones horizontales, tal como se indica por las flechas horizontales en la figura 1. Para aplicar el criterio de convergencia, se comienza con la comparación entre las variedades referentes, es decir, las variedades monolingües o menos bilingües (flecha horizontal superior). Esta comparación tiene como meta identificar diagnósticos de diferencia gramatical, que permitirán distinguir la influencia externa del cambio interno (o lo que se ha denominado puntos de conflicto estructural, *cfr*. Poplack & Meechan 1998: 132; Blas Arroyo 1999). Solamente tras precisar las diferencias entre dos lenguas se puede saber si un cambio en una la aproxima más a la otra cuando entran en contacto y, por lo tanto, si cabe la hipótesis de que el cambio en aquella es por la influencia de esta.

Con base en los diagnósticos de las diferencias entre las dos lenguas, se emprende la tarea principal, que es la comparación entre las variedades de las dos lenguas habladas por los mismos bilingües (flecha horizontal inferior). Esta operación se dirige a la pregunta de si los bilingües mantienen separadas sus dos lenguas, que a su vez exige al investigador determinar si las variedades en contacto muestran más similitud gramatical que sus referentes monolingües (o menos bilingües) y, lo que es primordial, establecer la dirección del cambio, es decir, si el cambio se encamina hacia una lengua o hacia la otra.

¿En qué criterios se basa la comparación? Las diferencias interlingüísticas pueden ser cualitativas —una estructura está presente en una lengua

y ausente en otra— o cuantitativas —una estructura es de uso preferido (o inclusive categórico) en una lengua y de uso minoritario en otra (Givón 1979: 22-43; Bresnan *et al.* 2001: 29; Torres Cacoullos & Travis 2019: 656). Aquí tomamos una perspectiva cuantitativa, enfocándonos en las estructuras variables, por las que entendemos, siguiendo a Labov (1969: 738), maneras gramaticales alternativas de «decir la misma cosa» («saying the same thing»), o formas morfosintácticas diferentes que cumplen funciones gramaticales generalmente similares. Nos enfocamos en las estructuras variables porque todo cambio presupone la variación (aunque no toda variación termina en cambio) (Weinreich *et al.* 1968: 188).

Como prueba cuantitativa de convergencia se ha apelado a las diferencias en la frecuencia global de una forma entre el referente monolingüe y la variedad bilingüe, por ejemplo, del sujeto pronominal versus al sujeto nulo en variedades bilingües del español en contacto con el inglés. Sin embargo, el cambio en el volumen de uso de una forma puede ser un criterio equívoco de la existencia del cambio lingüístico. Esto es porque el índice de frecuencia es susceptible de vicisitudes en la distribución de los datos (la abundancia o la escasez fortuitas de algún contexto altamente propicio en la base de datos), por cuestiones de género textual, la modalidad escrita u oral, el *priming* del interlocutor u otros motivos extragramaticales (Hernández 2009: 604-606; Poplack & Levey 2010: 404; Travis & Lindstrom 2016). Dadas las alteraciones por factores extralingüísticos, no se puede establecer con certeza el límite inferior para que una diferencia en el índice de frecuencia sea lingüísticamente significativa, de manera que sirva como criterio definitivo de diferencia gramatical.

El criterio cuantitativo, entonces, tiene que ir más allá del índice de frecuencia de la forma de interés, especialmente para las variables morfosintácticas que suelen ser sensibles a varios factores contextuales, factores que además pueden interactuar entre sí. Por ejemplo, la frecuencia de los sujetos nulos en inglés alcanza un índice cuatro veces mayor en narrativas que en la conversación en el contexto variable (40% *vs.* 10%), pero esta diferencia es consecuencia de la *distribución contextual* de los datos: en las narrativas hay mayor continuidad referencial (Travis & Lindstrom 2016: 112). La adecuación de las diferencias de frecuencia como criterio del cambio depende de la clase de variable y de la importancia y la complejidad de los condicionantes. Por ejemplo, para los cambios vocálicos una diferencia de frecuencia puede ser reveladora, ya que resulta más fácil controlar los condicionantes (por ejemplo, limitando el análisis a los contextos preobstruyentes). Pero para los pronombres sujetos, que son sensibles a varios factores además de la

continuidad referencial, hay que dar cuenta de la distribución contextual de los datos. Esto es precisamente lo que se logra mediante las comparaciones del condicionamiento lingüístico.

El condicionamiento lingüístico de las formas morfosintácticas que aparecen en alternancia proporciona un criterio más preciso de la diferencia gramatical. Por *condicionamiento lingüístico* se entiende un modelo de las restricciones probabilísticas sobre la forma morfosintáctica de interés, según su frecuencia en contextos lingüísticos particulares (Labov 1969) (*cfr*. Cedergren & Sankoff 1974). Volviendo al caso del sujeto pronominal, los numerosos estudios del español encuentran básicamente el mismo condicionamiento lingüístico sin importar la variedad y el índice de la frecuencia del pronombre. Por ejemplo, la restricción probabilística de la accesibilidad del referente opera en todas las variedades, de modo que el sujeto pronominal tiende a presentar una frecuencia superior en contextos no correferenciales que en contextos correferenciales. De la misma manera, en todas las variedades favorecen el sujeto pronominal las construcciones con verbos de cognición (p. ej. *(yo) creo*) y lo desfavorecen las formas verbales del pretérito indefinido. El condicionamiento lingüístico que comparten las variedades monolingües constituye el punto de referencia para establecer el cambio lingüístico en una variedad bilingüe, y también para establecer las diferencias interlingüísticas. Por lo tanto, los *diagnósticos cuantitativos de diferencia gramatical* entre dos lenguas en contacto se identifican tras la comparación entre los patrones de variación interna propios de cada una de las lenguas en contacto. Las diferencias en los patrones de variación sirven como criterio para diagnosticar la dirección del cambio en la variedad bilingüe y, por lo tanto, para evaluar si se trata de cambio debido al contacto.

3. Construcción de corpus bilingüe

En *Lenguas en contacto*, Weinreich hizo hincapié en que «los individuos hablantes son el *locus* del contacto» («the language-using individuals are thus the locus of the contact») (1968 [1953]: 1). Con el fin de estudiar de manera sistemática el habla de tales bilingües, abogamos por la creación de corpus fundados en cuatro principios: una comunidad de habla bien definida, muestras sistemáticas de miembros de la comunidad, grabaciones del habla espontánea y transcripción metódica. Solamente con un corpus construido sobre estas bases pueden contextualizarse los fenómenos de contacto, tanto en su marco social, porque permite estudiar los individuos en relación a los grupos a los que pertenecen, como en su marco lingüístico, porque un corpus del

habla íntegramente transcrito permite detectar cuáles fenómenos representan patrones sistemáticos y cuáles son de baja incidencia o meras idiosincrasias.

FIGURA 2. Los cuatro principios para un corpus sociolingüístico

Ilustramos aquí la aplicación de estos criterios por medio de la presentación del *Corpus bilingüe español-inglés de Nuevo México* (*New Mexico Spanish-English Bilingual (NMSEB) corpus*), basado en entrevistas sociolingüísticas realizadas en 2010-2011 (*cfr*. Torres Cacoullos & Travis 2018: cap. 2-4). Este corpus documenta la variedad del español hablada en el norte de Nuevo México, EEUU, junto con la variedad del inglés hablada por los mismos bilingües.

3.1. La comunidad de habla

Se ha argumentado a favor del contacto como motivo del cambio lingüístico apoyándose en fuentes de datos muy dispares, desde la publicación de ejemplos lingüísticos aislados, a los experimentos en laboratorios sicolingüísticos y a los corpus de habla. Los participantes que producen los datos, asimismo, han sido igual de heterogéneos: desde estudiantes universitarios aprendices de una segunda lengua hasta miembros de comunidades bilingües, por una parte, y desde comunidades en el proceso de desplazamiento de la lengua minoritaria hasta comunidades bilingües establecidas, por la otra.

Para el español de los EEUU los estudios se han enfocado en las comunidades de inmigrantes, por ejemplo, provenientes de México en Los Ángeles o de la República Dominicana en Nueva York. Una de las conclusiones ha sido que se trata de un bilingüismo social en el que se desarrolla un continuo de competencia oral, con desplazamiento hacia el inglés en los hablantes de las segundas y terceras generaciones (los hijos y los nietos de los inmigrantes) (Silva-Corvalán 1994: 11ss.). Numerosos estudios, además, han concluido que el contacto lingüístico da como resultado cambios en el español, en particular, aumentos de frecuencia de la variante con paralelo en el inglés, por ejemplo, la perífrasis *estar* + gerundio con respecto a la forma simple para expresar el aspecto progresivo, el modo indicativo en lugar del subjuntivo y

el sujeto pronominal versus el sujeto nulo (p. ej. Silva-Corvalán 1994; Otheguy & Zentella 2012). Sin embargo, la situación de un miembro de una comunidad bilingüe establecida, donde hay transmisión lingüística y los cambios no por fuerza son transitorios, se debe distinguir de la de un inmigrante de segunda o tercera generación que experimenta desplazamiento hacia el inglés o la de un estudiante universitario que aprende una segunda lengua.

Definimos una *comunidad de habla* como un grupo de hablantes que viven en una zona geográfica demarcada, participan en un conjunto de normas compartidas y tienen experiencias sociolingüísticas afines (*cfr.* Labov 1972: 120-121). Para una comunidad *bilingüe*, sus miembros comparten además la misma variedad de cada una de sus lenguas y convenciones unificadas para combinarlas. Una comunidad bilingüe *establecida* es aquella que mantiene suficiente continuidad para permitir el establecimiento de normas comunitarias y la transmisión de estas pautas tras más de una generación.

Circunscribir la comunidad de habla es imprescindible porque sabemos, gracias a los estudios comparativos de Poplack y sus colegas, que para los fenómenos bilingües «la habilidad personal de un individuo es regulada por las normas de su comunidad de habla» («an individual's personal ability [...] is mediated by the norms of his speech community») (Poplack *et al.* 1988: 98). Estas normas comunitarias pueden variar inclusive cuando las lenguas en contacto son las mismas. Un buen ejemplo de esto son las estrategias para incorporar verbos del inglés al español. Una estrategia es usar el verbo de apoyo *hacer* (p. ej. *lo hicieron hire* 'lo contrataron') (*cfr.* Wilson & Dumont 2015). Mientras que esta construcción es desconocida por los bilingües en Puerto Rico, es la opción de adaptación claramente preferida en Nuevo México, con más de la mitad de los casos en el corpus NMSEB, frente al uso de sufijos (*te baqueamos* 'te apoyamos', del inglés *back*) o de verbos auxiliares (*estaríamos discussing* 'estaríamos discutiendo'), N=104/195).

El norte de Nuevo México, la base del corpus que se presenta aquí, se colonizó, desde Nueva España, a fines del siglo XVI y conserva lo que podría ser la variedad más antigua de uso continuo del español en el continente americano que no ha sido actualizada por inmigración más reciente (Lipski 2008: 193) (*cfr.* Bills & Vigil 2008: 51-74). Hoy, de los estados norteamericanos, Nuevo México tiene la población hispana (o latina) más alta como porcentaje de la población total, 46% de sus dos millones de habitantes, seguido de 38% en California y en Texas (otros estados con cifras altas son la Florida, con 23%, y Nueva York con 18%). En algunos condados ('counties') de la parte norte del estado, el porcentaje de los hispanos alcanza el 80% (United States Census Bureau 2016).

En el norte de Nuevo México, se presenta una situación que, según propuestas que se han difundido acerca del contacto de lenguas, debería propiciar la convergencia gramatical, tanto por la poca interacción con hablantes de variedades monolingües de la lengua minoritaria como por el tiempo histórico largo de contacto con la lengua mayoritaria (p. ej. Thomason 2001: 65-76). En Nuevo México menos del 25% de los hispanos nacieron fuera de los EEUU y en partes del norte del estado, apenas un 5% (comparado con los estados de Texas, con 33%, y de Nueva York, con 39%) (United States Census Bureau 2016). Así que, a diferencia de los escenarios de contacto con las poblaciones inmigrantes que figuran en la mayoría de los estudios sobre la convergencia gramatical en los EEUU, en el norte de Nuevo México el contacto con hablantes monolingües de la lengua minoritaria está limitado.

Además, lleva una larga historia de contacto con la lengua mayoritaria. Por más de 150 años el contacto con el inglés ha sido intenso. En 1848 la región pasó de México a los EEUU al firmarse el Tratado de Guadalupe Hidalgo, en 1878 llegó el ferrocarril y más anglohablantes, en 1912 Nuevo México se hizo estado de los EEUU y ya para los años 40 se había impuesto el inglés en las escuelas públicas. El desplazamiento del español no se restringió al plan de estudios oficial. A los niños se les castigaba por hablar español en la escuela, tal como cuenta Pedro en (1), uno de los hablantes grabados para el corpus NMSEB, nacido en 1953. (En los ejemplos, la traducción aparece a la derecha, con los segmentos originalmente producidos en inglés en itálicas.)

(1) Sobre los castigos que se recibía en la escuela por hablar español

a.	*pero me acuerdo que,*	'pero me acuerdo que,
b.	*when we were in elementary,*	*cuando estábamos en la primaria,*
c.	*...(1.0) if you got caught uh=,*	*...(1.0) si te pillaban uh,*
d.	*.. s- --*	*..h- --*
e.	*uh speaking anything but English,*	*uh hablando otra cosa más que el inglés,*
f.	*...(1.1) uh=,*	*...(1.1) uh,*
g.	*you had to pay a price.*	*tenías que pagar un precio.*
h.	((8 líneas intercaladas))	
i.	*... usaban la jarita or you had to go out and get a load of wood.*	... usaban la jarita *o tenías que salir y traer una carga de leña.*'

[Pedro, 10 El Timbre Portátil: 10.59-11.18][2]

[2] Los ejemplos son extraídos del corpus NMSEB (Torres Cacoullos & Travis 2018: capítulos 2 & 3), a menos que se indique lo contrario. Entre paréntesis se da el nombre del hablante (pseudónimo), el número y el nombre de la transcripción, y el tiempo del comienzo y el fin del ejemplo en el archivo

Hoy en día ha vuelto el español a las escuelas, pero como lengua extranjera y en perjuicio del dialecto local, ya que se imparten variedades que se consideran «el español correcto», como lo describió uno de los participantes en el corpus. Otro hablante cuenta que su hija no quiso hablar español de niña pero que retomó la lengua cuando asistió a la universidad en el sur del estado entre compañeras de clase mexicanas. En el extracto en (2) califica de «bonito» el español de su hija, por no ser el español de la familia. Así, a pesar del porcentaje más elevado de hispanos de cualquier estado, el español nuevomexicano es un dialecto en peligro de extinción por la desvaloración frente a variedades monolingües, junto con el desplazamiento hacia el inglés (Bills & Vigil 2008: 313).

(2) La estimación del español mexicano en comparación con el nuevoméxicano (sobre su hija, quien había tenido compañeras de clase de México en la universidad)
   a. *lo hablaba muy bonito,*
   b. *como los de México.*
   c. *... porque .. aprendió más por ella,*
   d. *que por nosotros.*

[Trinidad, 21 Demerits: 03.33-03.39]

3.2. El muestreo de hablantes bilingües

El criterio de muestreo está determinado por la meta de la investigación. El habla de los miembros de la comunidad nuevomexicana que todavía mantienen su español proporciona datos valiosos para comprobar la hipótesis de la convergencia gramatical. Pero para poner a prueba si los bilingües tienen una gramática convergente, como se ha dado por supuesto en numerosas obras sobre lenguas en contacto, o si, en cambio, mantienen dos gramáticas independientes, primero hay que pormenorizar quiénes son los bilingües. Para el corpus NMSEB el criterio aplicado para ser bilingüe es el de hacer uso habitual en su vida diaria de ambas lenguas, según lo han comprobado los que realizan el trabajo de campo durante un periodo amplio de tiempo. Queda comprobado además el bilingüismo de los participantes por medio de su uso de ambas lenguas a lo largo de las grabaciones. En este punto también los participantes del corpus se diferencian de las poblaciones bilingües típicamente estudiadas, ya sea de inmigrantes, ya sea de aprendices de lenguas

de sonido. Los protocolos de la transcripción se presentan en el apéndice. Sobre la cuestión ética del anonimato de los entrevistados, véase Torres Cacoullos & Travis (2018: 47-48).

extranjeras en la universidad; mientras para estos grupos puede ser válido el constructo de primera versus segunda lengua o el de dominio lingüístico con base en pruebas formales, para los bilingües nuevomexicanos no lo es (Torres Cacoullos & Travis 2018: 62ss).

Los participantes cuya habla constituye el corpus NMSEB son 40 nuevomexicanos norteños de la cuarta generación mínimamente, o sea, que desde los abuelos, o más atrás, han estado en la zona. Representan mujeres (58%) y hombres (42%) que nacieron entre 1922 y 1993, mayormente residentes en áreas rurales (72%), y quienes cubren una gama de ocupaciones, que incluyen mineros, rancheros, maestros y empleados en varios servicios.

3.3. La entrevista sociolingüística

Para el análisis lingüístico, se privilegia el habla espontánea bilingüe debido a que se ha comprobado que «los datos más sistemáticos» provienen de la *lengua vernácula,* la producción no monitoreada del habla cotidiana (Labov 1972: 208). En estilos formales, además, los hablantes bilingües a menudo evitan precisamente los fenómenos que son de interés para el estudio del bilingüismo (Poplack 1981). Para aproximarse a la lengua vernácula, se ha desarrollado la técnica de la *entrevista sociolingüística* con el fin de grabar muestras del habla menos alterada por el automonitoreo o la hipercorrección (Labov 1984: 32-42).

La realización de las entrevistas por miembros de la comunidad puede además minimizar el impacto de un entrevistador de afuera, sea un académico asociado con la autoridad de la universidad o una persona no familiarizada con la realidad de la comunidad. En el caso de NMSEB, los entrevistadores fueron ocho estudiantes de la Universidad de Nuevo México, a quienes entrenamos en la entrevista sociolingüística. Hicieron grabaciones con un participante o con grupos de participantes (usualmente dos), mayormente miembros de su familia extendida o conocidos de ellos. Los entrevistadores no intentaron obtener piezas léxicas, estructuras gramaticales o fenómenos bilingües sino que se les pidió sencillamente que hablaran como lo harían de forma natural en español y en inglés.

El tipo de fenómenos bilingües que surgen de manera espontánea en las grabaciones se ejemplifica en (3). Hay incorporación al español de sustantivos de origen inglés como *los weekends* 'fines de semana' en la línea (e), así como inserción de marcadores de discurso como *you know* 'tú sabes' (d). También hay alternancia entre secuencias de más de una palabra en cada lengua, por ejemplo, *y yo era la única* que se produce yuxtapuesto a *I wanted to go to the nightclubs* 'yo quería ir a las discotecas' (a-b).

(3) Fenómenos bilingües en el habla espontánea

a.	Ivette:	*I wanted to go to the nightclubs y=,*	*'yo quería ir a las discotecas* y=,
b.		*.. yo era la única,*	.. yo era la única,
c.		*de todas las que íbamos,*	de todas las que íbamos,
d.		*... you know,*	*... tú sabes*,
e.		*que nos juntábamos en los weekends,*	que nos juntábamos en los fines de semana,
f.		*to go dancing,*	*para ir a bailar,*
g.		*or [whatever].*	*o [lo que sea].*
h.	Rafael:	*[mhm].*	*[mhm].*
i.	Ivette:	*yo era la única que no sabía arrear*.	yo era la única que no sabía arrear.' ((conducir un automóvil))

[Ivette, 06 El Túnico: 51.54-52.04]

El comportamiento lingüístico de las palabras aisladas suele diferenciarse del comportamiento de las secuencias: los sustantivos aislados procedentes del inglés que se producen en unidades prosódicas por lo demás completamente en español, como en (e), tienden a ser integrados en la morfosintaxis del español (Torres Cacoullos & Aaron 2003; Aaron 2015; Poplack 2018). Por otro lado, en la alternancia de secuencias de palabras, como en (a-b), cada secuencia tiende a mantener los patrones de la lengua a la cual pertenece. Es decir, la yuxtaposición de secuencias, pero no la incorporación de palabras aisladas, entraña alternancia entre dos sistemas gramaticales en la producción lingüística. Esta distinción es esencial porque, sin reconocer los distintos fenómenos bilingües, resulta imposible indagar en las consecuencias del bilingüismo.

El corpus NMSEB es un corpus bilingüe por excelencia ya que, debido a la alternancia entre secuencias del español y el inglés, las dos lenguas están igualmente representadas: de las más de 36 mil cláusulas que componen el corpus, aproximadamente la mitad son en español y la mitad en inglés (basado en la lengua del verbo principal) (Torres Cacoullos & Travis 2018: 67). Eso nos proporciona una representación suficiente para realizar análisis cuantitativos de las dos lenguas.

3.4. Transcripción metódica

Para convertir una colección de grabaciones en un corpus servible para el análisis lingüístico, hay que transcribirlas. La transcripción metódica sigue

protocolos precisos para producir un corpus estandarizado que permita análisis globales de diversos fenómenos.

El uso de ortografía estándar facilita la búsqueda de formas lingüísticas, asegura más consistencia entre transcriptores y además evita alentar percepciones negativas de la comunidad. Por ejemplo, en el corpus NMSEB no se representan variantes fonéticas como la aspiración de la /s/ (*loh muchachoh*), dejando la codificación de la realización del segmento para un análisis fonético sistemático. Por otra parte, sí se representan variantes morfológicas y léxicas, por ejemplo, formas como *vide* 'vi' y *muchito* 'muchacho' en Nuevo México.

Conviene realizar la transcripción de forma que quede alineado el texto con el audio (usando software de transcripción como el de ELAN, Lausberg & Sloetjes 2009). Esto asegura acceso fácil a la materia prima del corpus, lo cual mejora la calidad de la transcripción, agiliza el estudio de fenómenos de la fonética y facilita la corrección y el pulimento de la transcripción, incluyendo la corrección de los errores que se van encontrando durante los procesos de extracción de datos (lo cual es normal, sobre todo en los primeros años de uso del corpus).

La sintaxis en el discurso no es independiente de la prosodia y, por tanto, es deseable incluir en la transcripción información prosódica. Para realizar una transcripción prosódica del corpus NMSEB, aplicamos el método basado en la Unidad Entonativa ('Intonation Unit'), UE, «un tramo de habla producido bajo un contorno entonativo único y coherente» («a stretch of speech uttered under a single, coherent intonation contour») (Du Bois *et al.* 1993: 47). Para delimitar las UEs el transcriptor atiende a rasgos como el tono, que suele ser más alto al principio de una UE, y el ritmo, que es más lento al final de la UE (figura 3). Para captar esta información, cada UE se presenta en una sola línea, seguida de puntuación que indica la prosodia con que se termina. Por ejemplo, en (3) presentado anteriormente hay una serie de UEs con contorno entonativo continuativo (marcado con la coma), en las líneas (a-f), y se termina el ejemplo con contorno entonativo final (marcado con el punto), en (g). En la transcripción se señalan además algunas características que pueden representar dimensiones de la interacción o el esfuerzo cognitivo, como las pausas (con series de puntos, como en (b) y (d)), y el solapamiento (entre corchetes, como en (g) y (h)). (Las convenciones de transcripción están resumidas al final de este trabajo).

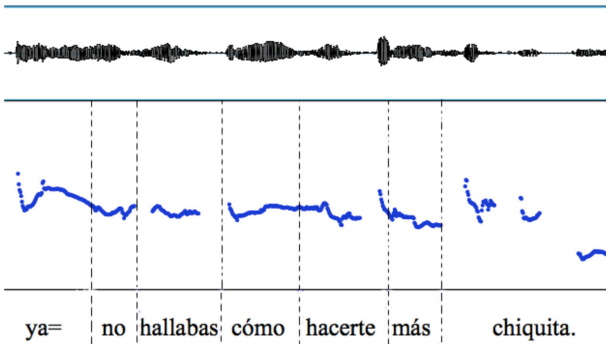


FIGURA 3. Propiedades de la Unidad Entonativa (UE) (*cfr.* Du Bois *et al.* 1993) (Torres Cacoullos & Travis 2018: 49)

La prosodia es relevante para el reconocimiento de las unidades en el discurso, porque la relación sintáctica tiende a ser más estrecha entre palabras producidas en una misma UE que entre palabras realizadas en UEs distintas (Croft 1995). Esta generalización nos ofrece, por ejemplo, el criterio de la separación prosódica para abordar la distinción entre sujeto y tópico o dislocación a la izquierda. Considérese una frase nominal como en (4) *la hija que vive aquí*, que se realiza sola en una UE que termina en un contorno entonativo de apelación, es decir, una subida de tono, marcado con el signo interrogativo al final de la unidad, que indica terminación y no continuidad, como la coma. Esta separación indica una relación con el verbo menos estrecha que la que existe entre la frase nominal sujeto y el verbo en la misma UE o en casos de contorno continuativo con una coma, como en (5), en que la frase nominal es claramente el sujeto, no obstante la intercalación de la oración de relativo. Se podría usar un corpus transcrito prosódicamente para investigar si la separación prosódica obedece a la complejidad de la frase nominal (por contener más de un modificador léxico, una frase preposicional o una oración de relativo) (Croft 2007: 19) o si sirve como medio presentativo que se tiende a aplicar a referentes no previamente mencionados.

(4) Separación prosódica entre frase nominal y verbo (con contorno entonativo de apelación)

... un= año para el aniversario,
*... **la hija que vive aquí?***
nos mandó una tarjeta.

[Mariana, 19 School Bus: 1:08:55-1:09:01]

(5) Conexión prosódica entre frase nominal y verbo (con contorno entonativo de continuación)

*pero **su hijo,***
que es troquero,
vive en California,

[Carlos, 26 La Pesca: 43:35-43:38]

4. Contacto y continuidad lingüística

Un corpus bien construido es un recurso permanente que se presta al estudio de diversas estructuras lingüísticas. Se ha hecho uso del corpus NMSEB, por ejemplo, para realizar desde estudios fonológicos (p. ej. Balukas & Koops 2015; Brown 2015; Zepeda 2018) hasta investigaciones sobre préstamos léxicos y alternancia de códigos (p. ej., Aaron 2015; Plaistowe 2015; Wilson & Dumont 2015; Steuck 2018) y análisis de variables sintácticas como la posición del sujeto en relación al verbo (Benevento & Dietrich 2015), el marcado diferencial del objeto (Sankoff *et al.* 2015) y la presencia del complementante (Steuck & Torres Cacoullos 2019). Aquí, dirigimos nuestra atención a la hipótesis de la convergencia gramatical.

Al cumplir con los cuatro criterios para la construcción de un corpus sociolingüístico resumidos en la figura 2, y además lograr representación de las dos lenguas en contacto, el corpus NMSEB permite enfrentar directamente la hipótesis de la convergencia gramatical: la producción de los hablantes bilingües, ¿responde a una sola gramática (convergente) o a dos gramáticas (independientes)? El criterio de convergencia convierte la hipótesis en una pregunta más precisa, a saber, si son más similares las gramáticas bilingües que las gramáticas monolingües, y exige comparaciones que nos permitirán responder esta pregunta de manera cuantitativa. Las comparaciones, a su vez, se apoyan en los diagnósticos cuantitativos de diferencia, basados en los patrones de variación interna.

4.1. Gramaticalización: las perífrasis progresivas

Uno de los mecanismos propuestos para el cambio atribuido al contacto es la *gramaticalización* de una estructura existente en la lengua minoritaria, que va adquiriendo el significado gramatical de una estructura paralela en la lengua mayoritaria (Heine & Kuteva 2005: 79-122). Por ejemplo, estructuras paralelas en el inglés y el español son las perífrasis aspectuales con gerundio que evolucionaron de una expresión locativa (Torres Cacoullos 2012). Se diferencian, sin embargo, en el grado de gramaticalización, ya que el proceso es más avanzado en inglés. En el inglés [*be* + verbo-*ing*] ha llegado a ser expo-

nente obligatorio del aspecto progresivo y la forma simple significa aspecto habitual (p. ej. *I'm drinking decaf* en este momento *vs. I drink decaf* habitualmente) (Bybee 2015: 193). En el español, por otro lado, la forma simple todavía alterna con [*estar* + Verbo-*ndo*] como expresión del aspecto progresivo (*tomo descafeinado* puede ser progresivo o habitual, según el contexto). En (6) y (7), por ejemplo, ambas formas expresan una situación vista como simultánea con el tiempo de referencia (en este caso, del verbo en pretérito indefinido en la cláusula anterior).

(6) Aspecto progresivo mediante la forma simple
(sobre alguien cuyo sombrero fue arrollado)
a. *cuando se le voló,*
b. *...(1.0)* ***venía*** *un carro y lo tropelló.*

[Rocío, 05 Las Tortillas: 18:59-19:03]

(7) Aspecto progresivo mediante la forma perifrástica
(sobre un piano que suena sin que nadie lo toque)
a. *y yo oí,*
b. *que* ***estaba tocando*** *el piano.*

[Rocío, 05 Las Tortillas: 28:54-28:56]

Esta variación en el español proporciona un diagnóstico cuantitativo de diferencia gramatical con el inglés. El cambio atribuido al contacto sería un aumento en el uso de [*estar* + verbo-*ndo*] respecto a la forma simple para expresar aspecto progresivo. Dumont & Wilson (2016: 410-412) comparan los índices de [*estar* + verbo-*ndo*] en el corpus NMSEB y en un referente monolingüe (un corpus comparable transcrito prosódicamente, Dumont 2016: 35-40), en dos contextos lingüísticos particulares: en contextos progresivos de pasado, como en (6) y (7), y en contextos habituales de pasado. En los dos contextos, el índice de [*estar* + verbo-*ndo*] es casi idéntico al del referente monolingüe: en contextos progresivos de pasado, 19% y 22% para el español nuevomexicano y el referente monolingüe respectivamente (N = 313 y 284 respectivamente); en contextos habituales de pasado, 3% y 4% (N = 509 y 91). No se observa, por lo tanto, una aproximación del español de los bilingües a los patrones del inglés, en contra de la hipótesis de aceleración de la gramaticalización por el contacto.

4.2. La propuesta de la simplificación: la variable subjuntivo - indicativo

Otro mecanismo propuesto para el cambio por el contacto es la *simplificación*, la erosión de una forma en la lengua minoritaria que no tiene forma análoga

en la lengua mayoritaria. Mientras que se ha reportado la simplificación en las situaciones de contacto transitorio, especialmente las de desplazamiento lingüístico (p. ej. Silva-Corvalán 1994: 3), la relevancia de la simplificación para una comunidad bilingüe establecida no ha sido comprobada.

Un ejemplo de la posible simplificación de una estructura del español en contacto con el inglés es el del subjuntivo en las oraciones subordinadas sustantivas. En el inglés el uso del subjuntivo ya no es productivo, mientras que en el español el mismo verbo matriz puede coaparecer con el subjuntivo o con el indicativo en la subordinada sustantiva, como sucede con *parecer* en (8) y (9). A favor de la simplificación se ha destacado la disminución del índice de la frecuencia del subjuntivo en la segunda y tercera generación respecto a la primera generación de inmigrantes en los EEUU.

(8) *Parecer* + subjuntivo
pero parece que ***pudieran*** *poner*$_{[SUBJ]}$ *a sign,* ((un aviso))
[Rubén, 29 La Diploma: 39:36-39:39]

(9) *Parecer* + indicativo
parecía que ***estaba***$_{[IND]}$ *toda detenida.*
[Ivette, 06 El Túnico: 36:50-36:52]

Como criterio más concluyente de la diferencia y por lo tanto del cambio gramatical, LaCasse (2018) aplica, primero, medidas de productividad del subjuntivo y, segundo, el criterio de los patrones de variación entre el subjuntivo y el indicativo en el corpus NMSEB y un referente español monolingüe (Martín Butragueño & Lastra 2011-2015).

En primer lugar, como indicio de la productividad del subjuntivo en los corpus, se toma en cuenta la distribución de los verbos matrices que aparecieron por lo menos una vez con una subordinada sustantiva que lleva el subjuntivo. Aproximadamente dos tercios de los verbos matrices aparecen solamente con el subjuntivo. Estos verbos matrices no variables en el español monolingüe se mantienen no variables en el español bilingüe (p. ej., *querer, dejar, pedir*), y en ambos dan cuenta de alrededor de la mitad de las apariciones del subjuntivo. Este par de resultados indica que la aparición del subjuntivo viene determinada en gran parte por el criterio léxico del verbo matriz, tanto en la variedad bilingüe como en el referente monolingüe.

Pero aun si el contacto no afecta al uso categórico, podría influir en el uso variable. Tal influencia se observaría en cambios en el condicionamiento con los verbos matrices variables, como *parecer* en (8) y (9) y *pensar* en (10) y

(11), por ejemplo, según los condicionantes de la polaridad y el tiempo verbal. Como muestra la figura 4, propician la elección del subjuntivo los contextos de polaridad negativa del verbo matriz como en (10) y (11) (panel a la izquierda), tanto en la variedad monolingüe como en la bilingüe. También son contextos favorables para el subjuntivo los de verbo matriz en forma del futuro, del imperativo, del condicional o del subjuntivo (panel a la derecha), en las dos variedades. No encontramos, entonces, evidencia de simplificación.

(10) *No pensar* + subjuntivo
yo no pensé que ***fuéramos*** *a salir/*
¿eh?/ yo pensé que íbamos a estar en la casa eh/ te comenté/ ¿no?
[CSCM 56, 108; Martín Butragueño & Lastra (2011-2015)]

(11) *No pensar* + indicativo
E: porque hasta la fecha/ uno piensa en un arquitecto/ en un hombre
I: sí
E: difícilmente en una mujer
I: sí
E: ¿no?
I: y no piensas/ que una señora/ que quiere una casa/ lo que necesita ***es*** *una mujer/*
[CSCM 31, 253-258; Martín Butragueño & Lastra (2011-2015)]

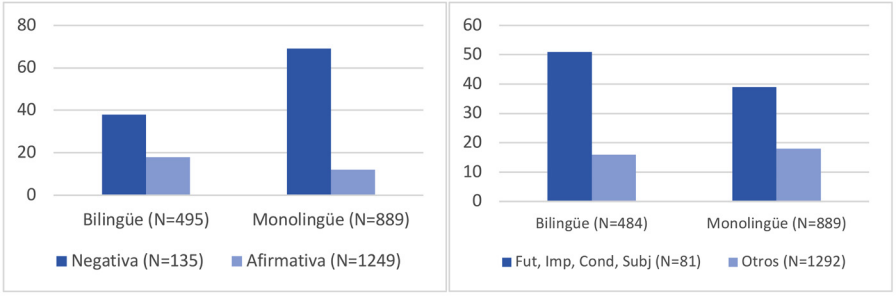


FIGURA 4. Índices del subjuntivo (*vs.* el indicativo) en subordinadas sustantivas, según la polaridad (cuadro a la izquierda) y el tiempo verbal (cuadro a la derecha) del verbo matriz, en el corpus nuevomexicano bilingüe (NMSEB) y en el referente monolingüe (de LaCasse 2018)[3]

[3] En la figura 4, el número de casos en el corpus bilingüe es un poco menor en el panel de la derecha que en el de la izquierda por no incluir verbos matrices del inglés, p. ej. *… I wished que, …(1.6) que alguien lo pudiera agarrar y kick him, 'deseaba* que alguien lo pudiera agarrar y *darle una patada'* (06, 06.46-06.51).

4.3. La hipótesis de la sobreextensión: expresión variable del sujeto pronominal en español y en inglés

El mecanismo de influencia externa más discutido es la *sobreextensión* de una forma variante que tiene una forma análoga en la otra lengua con frecuencia relativa más alta. Un caso ampliamente citado como caso de convergencia gramatical es la sobreextensión de los sujetos pronominales en una lengua de sujeto nulo en contacto con una lengua de sujeto no-nulo, que se manifiesta en un aumento en la frecuencia de uso (p. ej., Backus 2005: 333).

Este no es el caso para los bilingües nuevomexicanos. De acuerdo con las cifras de la figura 5, el índice del pronombre en el corpus NMSEB (en la segunda columna de cada figura) no es más alto del de otras variedades y cae entre los de Ciudad de México y Madrid.

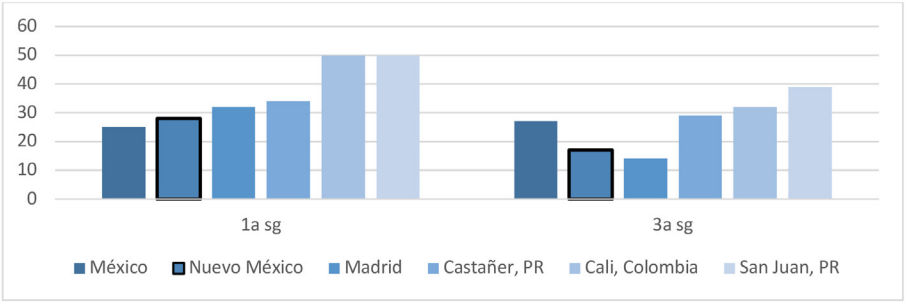


FIGURA 5. Índices de expresión del sujeto pronominal *vs.* Ø en variedades del español, para la 1ª sg y la 3ª sg. (de Travis & Torres Cacoullos 2019, figuras 3 y 4)[4]

Sin embargo, como ya adelantamos, el criterio de un índice «más alto» de sujetos pronominales es poco fiable. Un problema es cómo determinar lo que es un índice «alto», ya que las frecuencias globales de los sujetos pronominales varían de manera notoria entre las lenguas de sujeto nulo; por ejemplo, para la primera persona singular, 21% en el polaco frente a 48% en el portugués europeo (*cfr.* Torres Cacoullos & Travis 2018: 7). Además, dado que los índices pueden ser entre dos y tres veces más altos dentro del español monolingüe, según se ve en la figura 5, no hay base cierta para determinar lo

[4] Ciudad de México (Lastra & Martín Butragueño 2015: 43); Madrid, España (Enríquez 1984: 348); Castañer, Puerto Rico (Holmquist 2012: 208); Cali, Colombia (Travis & Torres Cacoullos 2018: 70); San Juan, Puerto Rico (Cameron 1992: 233).

que constituiría un aumento de frecuencia relevante —por estadísticamente significativo que sea— en una variedad bilingüe.

A pesar de las diferencias en el índice global, el condicionamiento lingüístico del sujeto pronominal actúa del mismo modo en las variedades del español. Es decir, al considerar la frecuencia de los sujetos pronominales en *contextos lingüísticos particulares*, las distintas variedades presentan resultados similares, por ejemplo, en los contextos no correferenciales y las construcciones de verbos de cognición, que propician el uso del sujeto pronominal, y en el contexto con el pretérito indefinido, que lo obstaculiza, como se mencionó anteriormente. Otra restricción probabilística es la persona gramatical. Según también se ve en la figura 5, la frecuencia de la variante pronominal tiende a ser más alta con los sujetos de la primera que con la tercera personal singular.

El examen del español de los bilingües nuevomexicanos revela que comparten el mismo condicionamiento lingüístico que se ha encontrado para hablantes de otras variedades del español (Torres Cacoullos & Travis 2018: 141-159). Pero aun así podría ser el caso de que haya convergencia si ha penetrado al español de los bilingües, siquiera de forma incipiente, alguna condición particular de la gramática del inglés. A continuación, aprovechamos los patrones de variación interna para identificar un diagnóstico cuantitativo de diferencia gramatical entre las dos lenguas.

Para este análisis, utilizamos corpus referentes monolingües comparables, transcritos prosódicamente siguiendo las mismas pautas que NMSEB (*cfr*. Travis 2005: 9-25) para el corpus referente del español y Du Bois *et al*. (2000-2005) para el inglés). Al basarnos en un corpus de habla espontánea podemos observar que, a pesar de la clasificación como lengua de sujeto no-nulo y, a pesar de la rareza relativa de los sujetos no expresados, en inglés también existe algún grado de variabilidad en la expresión del sujeto pronominal. Mostramos ejemplos de los referentes monolingües y del corpus bilingüe en (12) y (13) para el inglés y en (14) y (15) para el español. Indicamos los sujetos no expresados con un cero antes del verbo.

El análisis se centra en sujetos de la 1ª y 3ª persona singular con referente humano específico. Para el inglés, contamos todos los casos de sujeto no expresado (nulo), que alcanza frecuencias de no más del 2% para la 1ª persona y el 5% para la 3ª persona tanto en el corpus referente como en el corpus bilingüe (Torres Cacoullos & Travis 2018: 161). En vista de que los sujetos pronominales superan de lejos a los nulos, tomamos una muestra de aquellos, de manera que hubiera dos casos pronominales por cada sujeto no expresado (el procedimiento se describe en Torres Cacoullos & Travis 2018: 121-123).

(12) Expresión variable de sujeto en el inglés monolingüe

a.	*... and I put some onion powder ... in the mayonnaise,*	*'... y yo puse ajo en polvo ... en la mayonesa,*
b.	*.. and* **Ø** *put it on some .. boiled eggs.*	*.. y* **Ø** *la puse en unos .. huevos cocidos.*
c.	*...* **Ø** *Opened em up,*	*...* **Ø** *Los abrí,*
d.	*and I didn't stuff the eggs.*	*y yo no los rellené.*
e.	*I just put that mayonnaise on top.*	*Yo tan sólo puse esa mayonesa por encima.*

[Angela, SBCSAE 11: 759-763; Du Bois *et al.* (2000-2005)]

(13) Expresión variable de sujeto en el inglés bilingüe

a.	*.. I was able to scramble and,*	*'.. yo pude trepar y,*
b.	*.. find the --*	*.. hallar la --*
c.	*my other flashlight,*	*mi otra linterna,*
d.	*... Ø turned it on,*	*... Ø encendí esa,*
e.	*Ø worked on the one that I had just broke.*	*Ø me puse a reparar la que se acaba de romper.'*

[Manuel 16.1 Trip to Africa: 26.05-26.11]

(14) Expresión variable de sujeto en el español monolingüe
(sobre alguien en una foto que están mirando)

a.	Ángela:	*él es [muy lindo].*
b.	Dora:	*[Acá Ø tiene] barbi=ta,*
c.		*y acá no.*

[CCCS 24: 1712-1714; *cfr*. Travis (2005: 9-25)]

(15) Expresión variable de sujeto en el español bilingüe

a.	*...(1.0) I don't have that ... energy ya para hacer aquí en la casa como Ø tenía más antes.*	'...(1.0) *yo no tengo esa ... energía* ya para hacer aquí en la casa como Ø tenía más antes.
b.	*... porque yo venía del trabajo,*	... porque yo venía del trabajo,
c.	*...(1.5) and she would have the water you know,*	*...(1.5) y ella tenía el agua tú sabes,*
d.	*... ya soaking for the mud for .. the adobes?*	*ya remojando para el barro para .. los adobes?*'

[Miguel, 04 Piedras y Gallinas: 48:44-48:57]

La prosodia proporciona un diagnóstico cuantitativo de diferencia gramatical entre las dos lenguas. En el inglés, aparte de los verbos coordinados, como en (12) (a-b) y (c-d), la variabilidad está restringida a la posición prosódica inicial, como en (12) (c) y (e) y (13) (a), (d) y (e). En posición no inicial

de la Unidad Entonativa (UE), aparece un 100% de sujetos pronominales, es decir, nunca hay sujeto nulo. En el español, en cambio, no existe tal restricción, ya que alternan los sujetos pronominales con los nulos no solamente en posición entonativa inicial sino también en posición no inicial, como lo ilustran (a) y (b) en (15).

Sería posible, sin embargo, que el hecho de que los pronombres sujetos sean de aparición categórica en posiciones entonativas no iniciales en inglés influyera en el pronombre español, haciéndolo más frecuente en posición entonativa no inicial (por ejemplo, en cláusulas subordinadas). Como muestra la figura 6 (primer par de columnas), la tendencia cuantitativa en español va en dirección contraria a la restricción categórica del inglés: el índice del sujeto pronominal es más alto en posición inicial que en posición no inicial de la UE. La figura también indica que el uso del pronombre en posición no inicial no es más frecuente en el español nuevomexicano (tercer par de columnas), comparado con el corpus referencial. Es más, comparando los paneles de los referentes monolingües (a la izquierda) con los bilingües (a la derecha), se puede observar que las variedades bilingües siguen el patrón de sus respectivos referentes. El español de los bilingües no muestra una proclividad hacia la extensión de los pronombres en posición no inicial y tampoco en el inglés de los bilingües se erosiona la restricción contra los sujetos nulos en posición no inicial. La conclusión no puede ser otra que la de que no son más similares las variedades bilingües que los referentes monolingües entre sí.

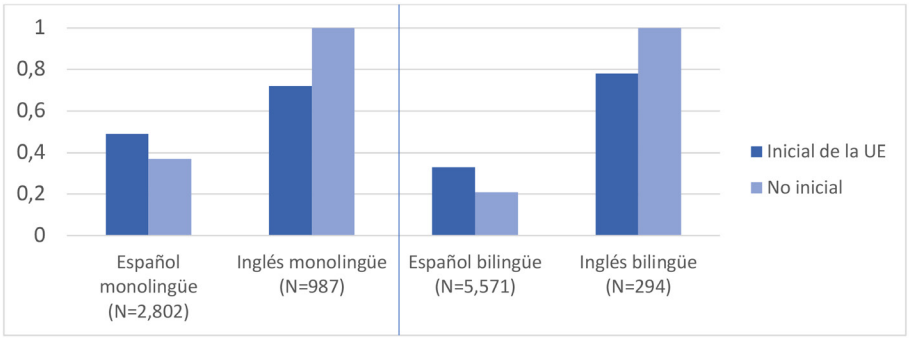


Figura 6. Índices de expresión del sujeto pronominal *vs.* Ø en el corpus nuevomexicano bilingüe (NMSEB) y en referentes monolingües, según la posición prosódica del sujeto en la Unidad Entonacional (UE) (de Torres Cacoullos & Travis 2018: 163)[5]

[5] El total de casos extraídos incluye los verbos coordinados, que no están representados en las columnas.

5. Conclusión

Hemos visto tres estudios fundamentados en la comparación del condicionamiento lingüístico de variantes morfosintácticas. Estos estudios del español nuevomexicano encuentran pruebas en contra de la convergencia gramatical y a favor de la continuidad lingüística, a pesar de que el escenario de contacto debería propiciar la convergencia. Los bilingües que cotidianamente practican sus dos lenguas pueden, entonces, mantener dos gramáticas independientes.

Bien se podría preguntar por qué no encuentran apoyo para la convergencia los análisis aquí resumidos, a diferencia de tantas afirmaciones previas sobre el contacto lingüístico. Una parte importante de la respuesta es la fuente de datos. En primer lugar, se trata de una comunidad bilingüe establecida y de participantes que mantienen el uso habitual de ambas lenguas, y no de una situación de contacto reciente por la inmigración, de participantes que no usan habitualmente la lengua minoritaria. Además, se apoya en un corpus del habla espontánea, tanto para las variedades bilingües como para los referentes monolingües o menos bilingües.

La otra parte importante de la respuesta es el método cuantitativo, basado en el condicionamiento lingüístico, ya que la frecuencia global de uso por sí sola no es criterio fiable de diferencia gramatical ni, por tanto, de cambio. La continuidad o el cambio lingüístico se establece tras comparar los patrones de variación interna, y especialmente según diagnósticos cuantitativos de diferencia gramatical.

Convenciones de transcripción (Du Bois *et al.* 1993)

.	contorno entonativo final	...	pausa mediana (aprox. 0.7 seg.)
,	contorno entonativo continuo	..	pausa corta (< 0.7 seg.)
?	contorno entonativo apelativo	...(N.0)	pausa larga, medida en segundos
--	contorno entonativo truncado	[ ]	habla simultánea, o solapamiento
=	sonido alargado	(( ))	comentario del investigador

Corpus

CCCS: Corpus of Conversational Colombian Spanish (*cfr.* Travis 2005: capítulo 2).

CSCM: Martín Butragueño, Pedro & Yolanda Lastra (2011-2015): *Corpus Sociolingüístico de la Ciudad de México*. Ciudad de México: El Colegio de México.

NMSEB: New Mexico Spanish-English Bilingual Corpus (*cfr*. Torres Cacoullos & Travis 2018: capítulos 2-4).

SBCSAE: John W. Du Bois, Wallace L. Chafe, Sandra A. Thompson, Charles Meyer & Robert Englebretson (eds.) (2000-2005): *Santa Barbara Corpus of Spoken American English*, Parts 14, Philadelphia: Linguistic Data Consortium.

Referencias bibliográficas

Aaron, Jessi Elana (2015): «Lone English-origin nouns in Spanish: the precedence of community norms», *International Journal of Bilingualism* 19/4, pp. 459-480. https://doi.org/10.1177/1367006913516021

Backus, Ad (2005): «Codeswitching and language change: one thing leads to another?», *International Journal of Bilingualism* 9/3-4, pp. 307-340. https://doi.org/10.1177/13670069050090030101

Balukas, Colleen & Christian Koops (2015): «Spanish-English bilingual VOT in spontaneous code-switching», *International Journal of Bilingualism* 19/4, pp. 423-443. https://doi.org/10.1177/1367006913516035

Benevento, Nicole & Amelia Dietrich (2015): «I think, therefore *digo yo*: variable position of the 1sg subject pronoun in New Mexican Spanish-English code-switching», *International Journal of Bilingualism* 19/4, pp. 407-422. https://doi.org/10.1177/1367006913516038

Bills, Garland D. & Neddy A. Vigil (2008): *The Spanish language of New Mexico and southern Colorado: a linguistic atlas*. Albuquerque, NM: University of New Mexico Press.

Blas Arroyo, José Luis (1999): «La gramática de la determinación en español y catalán: puntos de coincidencia y de conflicto estructural para la desambiguación de los fenómenos de contacto de lenguas», *Moenia: Revista lucense de lingüistica & literatura* 5, pp. 413-435.

Bresnan, Joan, Bresnan, Joan, Shipra Dingare, Christoper Manning & Miriam Butt (2001): «Soft constraints mirror hard constraints: voice and person in English and Lummi», in Miriam Butt & Tracy Hollaway (eds.): *Proceedings of the LFG01 Conference*. Stanford: CSLI Publications, pp. 13-31.

Brown, Esther L. (2015): «The role of discourse context frequency in phonological variation: A usage-based approach to bilingual speech production», *International Journal of Bilingualism* 19/4, pp. 387-406. https://doi.org/10.1177/1367006913516042

Bybee, Joan (2015): *Language change*. Cambridge: Cambridge University Press.

Cameron, Richard (1992): *Pronominal and null subject variation in Spanish: constraints, dialects, and functional compensation*. PhD thesis. University of Pennsylvania, Philadelphia, PA.

Cedergren, Henrietta & David Sankoff (1974): «Variable rules: performance as a statistical reflection of competence», *Language* 50, pp. 333-355. https://doi.org/10.2307/412441

Croft, William (1995): «Intonation units and grammatical structure», *Linguistics* 33, pp. 839-882. https://doi.org/10.1515/ling.1995.33.5.839

Croft, William (2007): «Intonation units and grammatical structure in Wardaman and in cross-linguistic perspective», *Australian Journal of Linguistics* 27/1, pp. 1-39. https://doi.org/10.1080/07268600601172934

Doğruöz, A. Seza & Ad Backus (2007): «Postverbal elements in immigrant Turkish: evidence of change?», *International Journal of Bilingualism* 11/2, pp. 185-220. https://doi.org/10.1177/13670069070110020301

Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming & Danae Paolino (1983): «Outline of discourse transcription», in Jane Edwards & Martin Lampert (eds.): *Talking data: transcription and coding in discourse*. Hillsdale: Lawrence Erlbaum Associates, pp. 45-89.

Dubert García, Francisco (2005): «Interferencias del castellano en el gallego popular», Bulletin of Hispanic Studies 83/8, pp. 271-291. https://doi.org/10.3828/bhs.82.3.1

Dumont, Jenny (2016): *Third person references: forms and functions in two spoken genres of Spanish*. Amsterdam: John Benjamins. https://doi.org/10.1075/sfsl.71

Dumont, Jenny & Damián Vergara Wilson (2016): «Using the variationist comparative method to examine the role of language contact in synthetic and periphrastic verbs in Spanish», *Spanish in Context* 13/3, pp. 394-419. https://doi.org/10.1075/sic.13.3.04dum

Enríquez, Emilia V. (1984): *El pronombre personal sujeto en la lengua española hablada en Madrid*. Madrid: CSIC, Instituto Miguel de Cervantes.

Givón, Talmy (1979): *On understanding grammar*. New York: Academic Press.

Goldrick, Matthew, Michael Putnam & Lara Schwarz (2016): «Coactivation in bilingual grammars: a computational account of code mixing», *Bilingualism: Language and Cognition* 19/5, pp. 857-876. https://doi.org/10.1017/S1366728915000802

Heine, Bernd & Tania Kuteva (2005): *Language contact and grammatical change*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511614132

Hernández, José Esteban (2009): «Measuring rates of word-final nasal velarization: the effect of dialect contact on in-group and out-group exchanges», *Journal of Sociolinguistics* 13/5, pp. 583-612. https://doi.org/10.1111/j.1467-9841.2009.00428.x

Holmquist, Jonathan (2012): «Frequency rates and constraints on subject personal pronoun expression: findings from the Puerto Rican highlands», *Language Variation and Change* 24/2, pp. 203-220. https://doi.org/10.1017/S0954394512000117

Labov, William (1969): «Contraction, deletion, and inherent variability of the English copula», *Language* 45/4, pp. 715-762. https://doi.org/10.2307/412333

Labov, William (1972): *Sociolinguistic patterns*. Oxford: Basil Blackwell.

Labov, William (1984): «Field methods of the project on linguistic change and variation», in John Baugh & Joel Sherzer (eds.), *Language in use: readings in sociolinguistics*. Englewood Cliffs, NJ: Prentice Hall, 28-53.

LaCasse, Dora (2018): *The subjunctive in New Mexican Spanish: maintenance in the face of language contact*. PhD thesis. Department of Spanish, Italian and Portuguese, Pennsylvania State University.

Lastra, Yolanda & Pedro Martín Butragueño (2015): «Subject pronoun expression in oral Mexican Spanish», in Ana M. Carvalho, Rafael Orozco & Naomi Lapidus Shin (eds.): *Subject pronoun expression in Spanish: a cross-dialectal perspective*. Washington DC: Georgetown University Press, pp. 39-57.

Lausberg, Hedda & Han Sloetjes (2009): «Coding gestural behavior with the NEUROGES-ELAN system», *Behavior Research Methods* 41/3, pp. 841-849. https://doi.org/10.3758/BRM.41.3.841

Lipski, John M (2008): *Varieties of Spanish in the United States*. Washington, DC: Georgetown University Press.

Otheguy, Ricardo & Ana Cecilia Zentella (2012): *Spanish in New York: language contact, dialect levelling, and structural continuity*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199737406.001.0001

Plaistowe, Jennifer (2015): *Coordinated code-switching? An investigation of language selection in bilingual conversation*. Honours thesis. Australian National University.

Poplack, Shana (1981): «Syntactic structure and social function of code-switching», in Richard P. Durán (ed.): *Latino language and communicative behavior*. Norwood, NJ: Ablex Publishing Corporation, pp. 169-184.

Poplack, Shana (2018): *Borrowing: loanwords in the speech community and in the grammar*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190256388.003.0004

Poplack, Shana & Stephen Levey (2010): «Contact-induced grammatical change: a cautionary tale», in Peter Auer & Jürgen Erich Schmidt (eds.): *Language and space: an international handbook of linguistic variation*, vol. 1: *Theories and methods*. Berlin: Mouton de Gruyter, pp. 391-419. https://doi.org/10.1515/9783110220278.391

Poplack, Shana & Marjory Meechan (1998): «Introduction: how languages fit together in codemixing», *International Journal of Bilingualism* 2/2, pp. 127-138. https://doi.org/10.1177/136700699800200201

Poplack, Shana, David Sankoff & Christopher Miller (1988): «The social correlates and linguistic processes of lexical borrowing and assimilation», *Linguistics* 26/1, pp. 47-104. https://doi.org/10.1515/ling.1988.26.1.47

Poplack, Shana, Lauren Zentz & Nathalie Dion (2012): «Phrase-final prepositions in Quebec French: an empirical study of contact, code-switching and resistance to convergence», *Bilingualism: Language and Cognition* 15/2, pp. 203-225. https://doi.org/10.1017/S1366728911000204

Rodríguez-Ordóñez, Itxaso (2017): «Reexamining differential object marking as a linguistic contact phenomenon in Gernika Basque», *Journal of Language Contact* 10, pp. 318-352. https://doi.org/10.1163/19552629-01002004

Sankoff, David, Nathalie Dion, Alex Brandts, Mayer Alvo, Sonia Balasch & Jackie Adams (2015): «Comparing variables in different corpora with context-based model-free variant probabilities», in Rena Torres Cacoullos, Nathalie Dion & André Lapierre (eds.): *Linguistic variation: confronting fact and theory*. New York: Routledge, pp. 335-346.

Silva-Corvalán, Carmen (1994): *Language contact and change: Spanish in Los Angeles*. Oxford: Clarendon Press.

Steuck, Jonathan (2018): *The prosodic-syntactic structure of intra-sentential multi-word code-switching in the New Mexico Spanish-English bilingual community*. PhD thesis. Department of Spanish, Italian and Portuguese, Pennsylvania State University.

Steuck, Jonathan & Rena Torres Cacoullos (2019): «Complementing in another language: prosody and code-switching», in Juan Andrés Villena Ponsoda, Francisco Díaz Montensinos, Antonio Manuel Ávila Muñoz & Matilde Vida Castro (eds.): *Language variation – European Perspectives VII*. Amsterdam: John Benjamins, 219-231.

Thomason, Sarah G. (2001): *Language contact: an introduction*. Washington DC: Georgetown University Press.

Thomason, Sarah G. & Terrence Kaufman (1988): *Language contact, creolization and genetic linguistics*. Berkeley, CA: University of California Press.

Torres Cacoullos, Rena (2012): «Grammaticalization through inherent variability: the development of a progressive in Spanish», *Studies in Language* 36/1, pp. 73-122. https://doi.org/10.1075/sl.36.1.03tor

Torres Cacoullos, Rena & Jessi Elana Aaron (2003): «Bare English-origin nouns in Spanish: rates, constraints and discourse functions», *Language Variation and Change* 15/3, pp. 289-328. https://doi.org/10.1017/S0954394503153021

Torres Cacoullos, Rena & Catherine E. Travis (2018): *Bilingualism in the community: code-switching and grammars in contact*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108235259

Torres Cacoullos, Rena & Catherine E. Travis (2019): «Variationist typology: shared probabilistic constraints across (non-)null subject languages», *Linguistics* 57/3, pp. 653-692. https://doi.org/10.1515/ling-2019-0011

Travis, Catherine E. (2005): *Discourse markers in Colombian Spanish: a study in polysemy*. Berlin: Mouton de Gruyter.

Travis, Catherine E. & Amy M. Lindstrom (2016): «Different registers, different grammars? Subject expression in English conversation and narrative», *Language Variation and Change* 28/1, pp. 103-128. https://doi.org/10.1017/S0954394515000174

Travis, Catherine E. & Rena Torres Cacoullos (2018): «Discovering structure: person and accessibility», in Naomi Lapidus Shin & Daniel Erker (eds.): *Questioning theoretical primitives in linguistic inquiry: papers in honor of Ricardo Otheguy*. Amsterdam: John Benjamins, pp. 67-90. https://doi.org/10.1075/sfsl.76.05tra

Travis, Catherine E. & Rena Torres Cacoullos (2020): «Beyond questionnaires: community-based measures of bilingualism», *Southwest Journal of Linguistics*, en prensa.

United States Census Bureau (2016): 2012-2016 5-Year American Community Survey. http://factfinder2.census.gov

Weinreich, Uriel (1968 [1953]): *Languages in contact: findings and problems*. The Hague: Mouton.

Weinreich, Uriel, William Labov & Marvin I. Herzog (1968): «Empirical foundations for a theory of language change», in Winfred P. Lehmann & Yakov Malkiel (eds.): *Directions for historical linguistics*. Austin, TX: University of Texas Press, pp. 95-195.

Wilson, Damián Vergara & Jenny Dumont (2015): «The emergent grammar of bilinguals: The Spanish verb *hacer* 'do' with a bare English infinitive», *International Journal of Bilingualism* 19/4, pp. 444-458. https://doi.org/10.1177/1367006913516047

Zepeda, Miguel (2018): *Debuccalization of /s/ and historic /f/ variation in Traditional New Mexican Spanish*. PhD thesis. Department of Spanish and Portuguese, University of California, Davis.

SOBRE LOS ORÍGENES DE LA CONSTRUCCIÓN ENCAPSULADORA EN ESPAÑOL

On the origins of the shell noun construction in Spanish

Anton Granvik
Universidad de Helsinki

Resumen

En este trabajo se analizan los orígenes de la construcción encapsuladora mediante una caracterización formal de esta construcción y de la función de «encapsulador nominal». La caracterización se basa en tres rasgos: la presencia de un artículo (definido o indefinido), la función sintáctica del sustantivo y el tipo de elemento (verbo, preposición, etc.) del que depende el sustantivo. Estos rasgos formales permiten clasificar los usos de los nueve sustantivos analizados —*causa, condición, convicción, esperanza, idea, noticia, ocasión, señal, sospecha*— como encapsuladores a) típicos, b) menos típicos, y c) marginales. Tras esta fase cuantitativa, basada en unos 7400 casos extraídos del *Corpus del Nuevo diccionario histórico del español*, se procede a su verificación cualitativa mediante un análisis diacrónico detallado que enfoca la función textual de los sustantivos de un número menor de casos, representativos de los tres estatus (usos típicos, menos típicos y marginales). Los resultados indican que los usos encapsuladores típicos, aunque son pocos, se documentan ya en la Edad Media y se van haciendo más frecuentes con el tiempo. También se observa que los usos típicos (según el análisis formal) normalmente tienen el formato N *es que* + cláusula en la época medieval, mientras que en los siglos más recientes se hacen más frecuentes los usos típicos de las construcciones N *de que* + cláusula y N *de* + infinitivo.

Palabras clave: construcción encapsuladora, oraciones completivas, sustantivos abstractos, sintaxis histórica, análisis de corpus

Abstract

This paper analyzes the origins of the shell noun construction in Spanish by means of characterizing the construction, and the function of shell noun, in terms of three formal

features: the presence of a determiner (definite or indefinite article), the syntactic function of the noun, and the kind of element (verb, preposition, etc.) on which the noun depends. These grammatical features allow classifying the uses of the nine analyzed nouns —*causa, condición, convicción, esperanza, idea, noticia, ocasión, señal, sospecha*— as a) typical, b) less typical, and c) marginal shell nouns. After the first, quantitative phase, based on roughly 7400 cases extracted from the *Corpus del Nuevo diccionario histórico del español*, the formal analysis is evaluated qualitatively by means of a detailed, diachronic analysis focusing on the textual functions of a more limited set of corpus cases, representative of the typical, less typical and marginal uses of the nouns. The results indicate that there are typical shell noun uses already in the medieval period, and that the typical shell uses become more frequent with time. It is also observed that the typical uses often appear in the constructional format N *es que* + clause in the medieval period, whereas from the 16th century on the typical uses are more often of the format N (*de*) *que* + finite clause or N *de* + infinitive.

Keywords: shell-noun construction, complement clauses, abstract nouns, historical syntax, corpus analysis

1. Introducción

Es bien sabido que los sustantivos abstractos, como *hecho, idea, prueba, cuestión*, etc., cumplen una función textual importante, actuando como encapsuladores de información más compleja que permiten al hablante manipular esa información de modo eficiente y sugerente[1]. Así, cuando algo se caracteriza como una *idea* o un *hecho*, por ejemplo, esto implica que el oyente o lector debe atribuirle a la información encapsulada en esos sustantivos abstractos un cierto estatus discursivo. Compárense estos dos sustantivos con otros como *disparate* o *rumor*, cuyas connotaciones semánticas son mucho más comprometedoras discursivamente. Así, aunque en (1) y (2) los sustantivos *idea* y *disparate* podrían intercambiarse, esto llevaría a interpretaciones claramente diferentes del mensaje transmitido originalmente, simplemente porque *disparate* implica una evaluación marcadamente negativa en comparación con *idea* (*cfr*. López Samaniego 2018: 131).

[1] Este trabajo se inserta en el Proyecto de investigación con referencia: FFI2015-64080-P del Ministerio de Economía y Competitividad: «Procesos de gramaticalización en la historia del español (V): gramaticalización, lexicalización y análisis del discurso desde una perspectiva histórica». Quiero expresar mis agradecimientos a la profesora María José Rodríguez Espiñeira por invitarme a participar en el Seminario *Corpus y Construcciones,* organizado en noviembre de 2018 en la Universidade de Santiago. Agradezco asimismo los comentarios oportunos que se hicieron a versiones previas de este trabajo, tanto de parte de los evaluadores anónimos como de parte de las editoras de este Anexo de *Verba*, pues han contribuido significativamente a su mejora. Todos los errores e incongurencias restantes son de mi entera responsabilidad.

(1) Claro que me preocupa esa pobre gente. No me gusta *la idea de* empezar la «práctica» diseñando construcciones que van a desalojar a casi 5 mil almas, como dicen los curas... (CDH, 1992)

(2) ¿De dónde provenía *el disparate de* que dudara de mi origen? ¿No repetía que Victorien Sardou y ella habían estudiado cuanto concierne a la historia y el arte de mi país? ¡Ah, si yo hubiera podido escribir! (CDH, 1982)

Ahora bien, aunque en la actualidad estos usos parecen de lo más normales y corrientes —de hecho, en palabras del lingüista alemán Hans-Jörg Schmid (2000: 13) «one would not get along in discourse if it were not for the encapsulating function of shell nouns»—, en la lengua medieval es imposible encontrar ejemplos como los de (1) y (2). *Idea*, por ejemplo, no se usa en este tipo de función hasta el siglo XVIII. En el *Corpus del Nuevo diccionario histórico del español* (CDH), este sustantivo aparece por primera vez a principios del siglo XV, pero hay que esperar hasta el siglo XVIII para encontrar casos en los que *idea* se combina con oraciones completivas, cumpliendo la función de encapsular la información contenida en la completiva. En la misma línea, hablando de la función encapsuladora de los pronombres demostrativos neutros (*esto, eso, aquello*, principalmente), Borreguero (2018: 184, n. 8) señala que los encapsuladores anafóricos «constituyen un mecanismo de cohesión textual casi inexistente en español con anterioridad al s. XVII».

Sobre esta base, lo que me propongo hacer en este trabajo es intentar dar cuenta de los orígenes de la «construcción encapsuladora» en español. Entre las preguntas concretas que intentaré contestar se encuentran las siguientes:

— ¿Puede hablarse, en realidad, de una construcción encapsuladora en los datos medievales, o el uso encapsulador es una innovación más bien del llamado español clásico o moderno?
— ¿Cuáles son los primeros sustantivos que funcionan como encapsuladores? ¿Qué función textual tienen?
— ¿Cómo se relaciona la construcción gramatical con la función textual, por un lado, y con los diferentes sustantivos y épocas, por otro?

Como revelan estas preguntas, un aspecto importante a la hora de investigar lo que en sentido general puede llamarse la «encapsulación nominal» lo constituye la construcción gramatical en la que se insertan los sustantivos abstractos: N + oración completiva. Este aspecto se tratará en el apartado 2, donde se presenta un repaso teórico de la construcción encapsuladora.

El apartado 3 está dedicado al método y al corpus empleado, mientras que el análisis propiamente dicho se presenta en el apartado 4. El trabajo concluye con una discusión de los resultados y unas consideraciones finales en el apartado 5.

2. Las bases - la encapsulación nominal, la construcción encapsuladora y las oraciones completivas de sustantivo

La bibliografía especializada describe la *encapsulación nominal* como «el empleo de SSNN definidos cuyos sustantivos, que actúan como núcleos del sintagma, tienen como antecedente anafórico —o catafórico, siendo en este caso, *consecuente* o *poscedente*— a un fragmento textual que puede ser de diversa extensión y complejidad conceptual» (Abad Serna 2015: 229). Esto se observa en el ejemplo (3):

(3) En las últimas décadas ha tenido lugar un cambio en la concepción tradicional de la ciencia. Ha surgido una nueva filosofía de la ciencia; en *esta nueva concepción* se rechaza *la idea* de que puede haber observaciones teóricamente neutrales [...] (CDH, 2003)

El ejemplo (3) incluye, en realidad, dos encapsuladores nominales, el sustantivo *concepción*, que hace referencia anafórica al segmento textual inmediatamente anterior —el «cambio en la concepción tradicional» y la «nueva filosofía de la ciencia»—, y el sustantivo *idea* que encapsula (a modo de catáfora) la oración completiva que le sigue.

Nótese que ni la construcción nominal *esta nueva concepción* ni *la idea de que* son realmente necesarias desde un punto de vista semántico. Por ejemplo, podríamos decir, simplemente, «ha surgido una nueva filosofía de la ciencia, en la que se rechaza la idea...» y el mensaje sería esencialmente el mismo. Asimismo, si dijéramos «en esta nueva concepción se rechaza que pueda haber observaciones teóricamente neutrales», tampoco se alteraría mucho el significado transmitido. Lo que añade la presencia de un encapsulador como *idea*, en este caso, es una caracterización de la proposición «que puede haber observaciones teóricamente neutrales» como una idea, es decir, como un tipo de entidad mental. La función de *idea* en (3) se corresponde perfectamente con la descripción de Abad Serna (2015: 230) según la cual «los sustantivos empleados como encapsuladores tienen la singularidad de *sintetizar* la referencia del segmento textual al que remiten».

Un rasgo gramatical importante mencionado por Abad Serna (2015) es que «los encapsuladores constituyen SSNN plenos con referencia autónoma»

(*cfr*. González Ruiz 2009: 252, *apud* Abad Serna 2015: 232), donde con «SSNN plenos con referencia autónoma» se refiere a «sintagmas constituidos por elementos que no son proformas pronominales, y que tienen referencia propia» (Abad Serna 2015: 177). Algo que se observa tanto con respecto a *esta nueva concepción* como con respecto a *la idea* en el ejemplo (3).

Hablando de los sustantivos encapsuladores (*shell nouns*, en inglés), Schmid (2000, 2018) observa que el estatus de encapsulador es una propiedad funcional, lo que quiere decir que «un sustantivo dado se convierte en encapsulador cuando un hablante decide usarlo en una construcción encapsuladora con fines específicos» (2000: 13)[2]. Para Schmid (2000: 14) la construcción encapsuladora funciona en tres dimensiones diferentes, una semántica, una cognitiva y una textual:

> Función semántica[3]
>
> «Characterizing and perspectivizing complex chunks of information.»
>
> Función cognitiva
>
> «Temporary concept-formation (encapsulate these complex chunks … in temporary nominal concepts).»
>
> Función textual
>
> «Linking these nominal concepts with clauses.»

Según Schmid (2000: 21), la función encapsuladora puede considerarse una categoría prototípica, cuyo núcleo es la relación de identidad experiencial (*experiential identity*, Schmid 2000: 21; *cfr*. López Samaniego 2011: 447; Abad Serna 2015), es decir, «the notion that the shell noun and the shell content express ideas about the same thing.» Volviendo al ejemplo (3), arriba, esta identidad experiencial puede verificarse al constatar que la oración «que puede haber observaciones teóricamente neutrales» es una *idea*. Ahora bien, si el núcleo de la categoría es la noción de identidad experiencial, es de esperar que a su alrededor existan casos menos típicos e, incluso, marginales. Así, para Schmid (2000: 25-26), los sustantivos usados como parte de «predicados expandidos», en combinación con verbos funcionales del tipo *tener* y *hacer,* «no pueden considerarse buenos ejemplos de sustantivos encapsuladores». Asimismo, los sustantivos de significado temporal o locativo, como

[2] La traducción es mía: el original en inglés lee: «A noun is turned into a shell noun when a speaker decides to use it in a shell-content complex in the service of certain aims.»

[3] Para los propósitos de este trabajo, que tiene una orientación diacrónica descriptiva, no parece muy importante la distinción entre las funciones semántica y cognitiva, por lo que en el análisis las trataré como un todo y hablaré de funciones semántico-cognitivas (véase el § 4 de este trabajo).

momento, lugar… «son tratados como marginales» (Schmid 2000: 26). Los ejemplos (4), (5) y (6) ilustran tres grados de tipicidad, lo cual se observa en que la noción de «identidad experiencial» resulta cada vez más difícil de sostener:

(4) Finalmente, pasaron dos años de este suceso, al cabo de los cuales Lisardo, consolado, que el tiempo puede mucho, salía en los calores de un ardiente verano a bañarse al río. Súpolo Marcelo, que siempre le seguía, y desnudándose, una noche, fue nadando hacia donde él estaba y le asió tan fuertemente, que con la turbación y el agua perdió el sentido y quedó ahogado, donde con gran dolor de toda la ciudad le descubrió la mañana en las riberas del río. Esta fue la prudente venganza, si alguna puede tener este nombre: no escrita, como he dicho, para ejemplo de los agraviados, sino para escarmiento de los que agravian, y porque se vea cuán verdadero salió *el adagio de que* los ofendidos escriben en mármol y en agua los que ofenden, pues Marcelo tenía en el corazón la ofensa, mármol en dureza, dos largos años, y Lisardo tan escrita en el agua, que murió en ella. (Lope de Vega, *La prudente venganza,* 1620, *apud* CORDE)

(5) Salen el REY, el CONDESTABLE y gente, don RODRIGO, y don FERNANDO
LEONOR ¡El rey!
PEDRO Llegad a besar su mano.
INÉS ¡Qué alegre llego!
PEDRO Dé vuestra alteza los pies, por la merced que me ha hecho del alcaidía de Burgos, a mí y a mis hijas.
REY *Tengo bastante satisfacción* de vuestro valor, don Pedro, y *de que* me habéis servido.
PEDRO Por lo menos lo deseo.
REY ¿Sois casadas?
INÉS No, señor.
REY ¿Vuestro nombre?
INÉS Inés.
REY ¿Y el vuestro?
LEONOR Leonor.
REY En gallardos caballeros emplearéis vuestras dos hijas…
(Lope de Vega, *El caballero de Olmedo*, 1610 *apud* CORDE)

(6) Conde Astolfo Loco soy, no dizes mal, y de la Reyna también.
Rey Teodosio ¿Quieres que *albricias* te den?
Conde Astolfo Sí, *de que* ya estoy mortal.
Rey Teodosio Su Magestad ¿cómo viene?
Conde Astolfo Con mucho disgusto mío.
(Lope de Vega, *El amigo por fuerza*, c 1599, *apud* CORDE)

Así, en (4) es bastante claro que la oración «los ofendidos escriben en mármol…» constituye un *adagio*. En (5), en cambio, aunque es posible que la *satisfacción* consista en «que me habéis servido», este hecho también es el motivo de la satisfacción, por lo que la preposición *de* puede interpretarse como marca de una relación causal (y su presencia tendría, por tanto, motivación semántica). Finalmente, en (6) no parece haber ni relación de identidad ni interpretación causal, entre *albricias* y «que ya estoy mortal» – pero sí una dependencia sintáctica (*albricias* rige *de* con este tipo de complementos); y el matiz semántico presente tiene más de motivo, causa que de relación de identidad (*cfr*. Granvik 2012: 251, 272-275).

Siguiendo la caracterización de Schmid (2000), a la hora de determinar si un uso determinado de un sustantivo abstracto constituye, o no, un caso de encapsulación nominal, lo natural sería tomar como criterio definitorio la noción de identidad experiencial. Es decir, se podría considerar que un nombre tiene función encapsuladora siempre y cuando hay «identidad experiencial» entre el sustantivo y algún elemento presente en el contexto inmediato (típicamente, la oración completiva). Así, si al aplicar el test de ¿el N es igual a la oración? (o, ¿la oración es (un/a) N?) la respuesta es afirmativa, se trataría de un uso encapsulador. En palabras de López Samaniego (2011: 446-447), «la actualización discursiva del nombre *solo* aparece en la cláusula completiva cuando esta mantiene una relación apositiva de identidad con el nombre»[4]. La alternativa es que la cláusula completiva tenga función de predicado, constituyendo un argumento del núcleo nominal, lo cual refleja la distinción entre completivas argumentales y apositivas de Leonetti (1993, 1999).

Sin embargo, la distinción entre identidad experiencial y la falta de ella no siempre es muy clara (y menos, objetiva). No hay, por ejemplo, una relación biunívoca entre una o más construcciones sintácticas (por ejemplo, N *de que*) y esta relación de identidad. Como revela el ejemplo que discute López Samaniego (2011: 446), tomado de Leonetti (1999), (6a) *La explicación [de que suspendas siempre] es que no te preocupas por entender bien la asignatura*, *vs*. (6b) *La explicación [de que han aumentado los gastos] no es muy convincente*, un mismo sustantivo puede combinarse con completivas tanto argumentales (6a) como apositivas (6b) en la misma construcción sintáctica.

En los datos diacrónicos que manejo en este trabajo, ocurre algo parecido. Con algunos sustantivos, como *esperanza*, por ejemplo, aunque casi siempre es posible pensar que hay identidad experiencial entre el sustantivo y la completiva, a veces esta identidad no resulta muy intuitiva, como ocu-

[4] Las cursivas son mías.

rre en el ejemplo (7); con los sustantivos *idea* y *noticia(s)*, en (8), (9) y (10), se puede observar una situación parecida. Así, en (8) aunque es posible que la *idea* sea lo mismo que *haberlas visto*, parece más natural pensar que se tiene esa idea 'acerca de' *haberlas visto*, asignándole a la preposición *de* el significado de tema/asunto (*cfr*. Granvik 2014). La misma noción de 'acerca de' parece bastante intuitiva también en los dos casos con *noticia(s)* en (9) y (10). Más que decir que el *haber llegado* constituye las *noticias*, parece que las noticias tratan de ello.

(7) Horrebow, Jacobo Cassini y algún otro tuvieron también alguna lisonjera *esperanza de haber* hecho este descubrimiento; pero fue infundado su juicio (CDH, 1793)

(8) Las hojas como las del «sanguiño», o de la madreselva. Tengo confusa *idea de haberlas visto* en otra parte y de que los llamaron «jazmines reales». (CDH, 1745)

(9) Dispónese a marchar a Vera Cruz con ánimo de esperar las órdenes de la corte y se halla con *noticias de haber llegado* a las costas algunos navíos españoles con tropas mandadas por Pánfilo de Narváez, cuyo objeto era prenderle. (CDH, 1774)

(10) y a principios del año de treinta y seis llegaron a vn parage de tierra tan estéril y empollada, llena de tantas ramblas y quebradas, que le pusieron por nombre el Mal-País, de cuyos habitadores (que eran pocos) huvieron algunos a las manos, que respondieron más por señas que con palabras a lo que les preguntaron, dieron alguna *noticia de que* a poca distancia de aquel sitio, azia la mano izquierda, estaba vn pueblo de vecindad quantiosa (CDH, 1723)

Un caso distinto, pero igualmente complicado desde la perspectiva de una caracterización semántica de la relación entre sustantivo y completiva, lo ofrecen los sustantivos *causa* y *señal*. Una oración completiva finita o de infinitivo del sustantivo *causa* normalmente no puede constituir la causa en sí misma, sino que corresponde más bien a lo causado, el efecto (igual que ocurre con la cláusula [*de que suspendas siempre*] del ejemplo (6a) de Leonetti, mencionado arriba). Esta situación se ilustra en (11), donde la completiva indica lo causado, y se interroga sobre cuál es la causa de ello. *Señal*, en el ejemplo (12), ilustra un caso semejante. La oración introducida por *que* no es la *señal*, sino lo señalado; pero la señal sí se identifica contextualmente, *las lágrimas llegan en aquella sazón*, en este caso.

(11) Preguntávanle al Diógenes que qué era *la causa que* los hombres davan más fácilmente la limosna a los coxos y mancos pobres (CDH, 1540)

(12) mucha tentaçión trae desesperaçión; mas si lágrimas llegan en aquella sazón, *señal es que* misericordia de Dios e consolaçión nos enbía, por que çiertos seamos de saluaçión (CDH, 1378; copia de mediados del s. xv))

Es decir, ni con *causa* ni con *señal* puede hablarse, en muchos casos, de identidad experiencial entre sustantivo y completiva. Por eso, si el criterio para identificar casos encapsuladores fuera la identidad experiencial, estos ejemplos no se calificarían como tales. Aun así, en ambos ejemplos los sustantivos funcionan como encapsuladores en el sentido de constituir nexos entre dos elementos discursivos, por lo que quizás no deberían descalificarse.

Dado que la relación de identidad entre el sustantivo y la oración completiva, por un lado, puede resultar difícil de establecer de modo objetivo[5] y, por otro, puede no ser necesario para que un sustantivo funcione como encapsulador en el sentido de hacer de nexo entre dos elementos discursivos, he optado por un acercamiento alternativo —formal y, por tanto, objetivo— al análisis de los sustantivos encapsuladores. Partiendo de las observaciones de investigadores anteriores (Schmid 2000, Rodríguez Espiñeira 2015), que señalan que son fundamentalmente cuatro las construcciones gramaticales que «convierten» un sustantivo en encapsulador:

a) demostrativo + nombre (por ejemplo, *este dilema*…)
b) nombre seguido de cláusula modificadora (*el dilema de que/* infinitivo…)
c) nombre como sujeto de una cláusula de identificación con el verbo *ser* (*el dilema es que*…)
d) demostrativo neutro o pronombre personal neutro como sujeto y nombre encapsulador como predicado (*esto es un dilema*) (según Rodríguez Espiñeira 2015: 658; los ejemplos son míos),

considero que todas las combinaciones de N con oración completiva (del formato N *de que*, N *que*, N *de* + infinitivo y N *ser que*/inf., es decir, las construcciones incluidas en los puntos b) y c), constituyen un caso de la construcción encapsuladora. Esta decisión encuentra apoyo en la consideración de Schmid (2018: 115) de que estos dos patrones léxico-sintácticos (N *ser* + oración

[5] Con datos diacrónicos, la interpretación objetiva de casos extraídos de un corpus supone un reto aun mayor.

y N *de* + oración) «destacan sobre los demás por constituir el tipo de contexto en el que los sustantivos encapsuladores triunfan» [6]. Entiendo, pues, que existe una construcción encapsuladora en el sentido esquemático que le otorga la Gramática de construcciones (*cfr*. Croft & Cruse 2008 [2004]; Gonzálvez-García 2012; Traugott & Trousdale 2013), en la que entran cuatro variantes formales: N *que* + oración, N *de que* + oración, N *de* + infinitivo y N *ser* + oración (finita o de infinitivo) (*cfr*. la discusión sobre el estatus de construcción de la construcción encapsuladora en Schmid 2018: 124-125).

Más allá de las cuatro estructuras sintácticas que acabo de definir, me apoyaré en tres criterios formales adicionales que figuran en las caracterizaciones de los nombres encapsuladores en la bibliografía y que me han servido para operacionalizar los usos encapsuladores:

1) La presencia de determinante (artículo, posesivo, demostrativo)[7]
2) Uso con verbo no ligero/de apoyo
3) La función sintáctica del N (sujeto, objeto directo, complemento preposicional, predicado)

Aparte de que los criterios formales que acabo de establecer me ayudan a identificar y caracterizar los usos encapsuladores, al partir de un macrocorpus como es el CDH, el no tener que basarme en análisis minuciosos del contexto sintáctico de cada concordancia —que muchas veces no incluye el contexto suficiente para realizar ese análisis cualitativo necesario—, hace que sea mucho más fácil recuperar los ejemplos relevantes.

Como explicaré con más detalle en el apartado siguiente (§ 3 Córpus y método), sobre la base de los tres rasgos formales característicos del contexto de uso de los casos extraídos del corpus, establezco una tipología de la construcción encapsuladora, según el uso sea formalmente típico, menos típico o marginal. Aunque no pretendo que esta caracterización basada exclusivamente en criterios formales sea un modo inequívoco de identificar los usos encapsuladores y su función discursiva-textual, sí intentaré mostar, con ayuda de un análisis contextual detallado, que la estructura formal típica supone un buen punto de partida para el análisis de su función discursivo-textual.

[6] La traducción es mía. El original lee «patterns a. and b. stand out from the others by providing the kind of environment in which shell nouns thrive».

[7] Nótese que el hecho de considerar solo la presencia *vs.* ausencia de determinantes no implica que la función del posesivo y del demostrativo se considere en todo equiparable a la del artículo. Sí implica, en cambio, que los grupos nominales en cuestión puedan considerarse «SSNN plenos con referencia autónoma» (*cfr*. Abad Serna 2015: 232).

3. Corpus y método

El corpus empleado es el *Corpus del Nuevo diccionario histórico del español* (CDH), más exactamente, su llamado *corpus nuclear*, que «cuenta con más de 62 millones de ocurrencias, de las cuales 38 pertenecen a textos españoles y más de 24 millones a obras americanas» (CDH, *Ayuda*). He decidido usar este corpus puesto que incluye información detallada acerca de la datación de los documentos medievales incluidos, distinguiendo entre la (supuesta) fecha de publicación de las obras y la fecha de las copias conservadas (*cfr*. Kabatek 2016; Rodríguez Molina & Octavio de Toledo 2018).

De este corpus he extraído las concordancias de nueve sustantivos diferentes que figuran entre los más frecuentemente usados en diferentes épocas de la lengua (*cfr*. Granvik 2018). Los nueve sustantivos son *convicción, esperanza, idea* y *sospecha* (sustantivos mentales, según la clasificación de Schmid 2000), *señal* y *causa* (factuales, según Schmid 2000) así como *condición* (circunstancial), *noticia* (lingüístico), *ocasión* (modal). Así, la muestra incluye sustantivos de las cuatro clases más importantes de Schmid (2000, 2018), más uno circunstancial. Entre los nueve sustantivos, *causa, condición, esperanza* y *sospecha* son típicos de la época medieval; *noticia, ocasión* y *señal* son característicos de la época clásica (siglos XVI y XVII), mientras que *convicción* e *idea* son encapsuladores típicamente actuales (*cfr*. Granvik 2017b: 249).

Las concordancias se realizaron enfocando tres/cuatro construcciones diferentes[8]:

— N *de* + infinitivo
— N *que/de que* + cláusula
— N *ser que* + cláusula

Las búsquedas se hicieron usando las siguientes fórmulas: N *de* (proximidad entre N y *de* uno a la derecha), N *que* (proximidad uno a la derecha) y N *ser* (proximidad uno a la derecha). El número total de concordancias extraídas es de 27555; estas fueron revisadas manualmente tras lo que se pudo establecer una muestra final de estructuras completivas que comprende unos 7400 casos.

A continuación, los 7400 casos fueron etiquetados según tres criterios formales, a saber: 1) si el sustantivo va o no acompañado de un determi-

[8] Aunque tanto Schmid (2000) como Rodríguez Espiñeira (2018) tratan las completivas finitas (N *de que*) y de infinitivo como pertenecientes al mismo tipo (nombre seguido de cláusula modificadora), en Granvik (2017a, 2017b, 2018) se demuestra que hay diferencias entre las completivas finitas y las de infinitivo (2017a) así como entre las introducidas por *de que* y las encabezadas por un simple *que* (2017b, 2018).

nante; 2) el núcleo sintáctico del sustantivo (verbo, verbo de apoyo, preposición, etc.), 3) la función sintáctica del sustantivo en la oración en la que se inserta (sujeto, objeto directo, complemento preposicional, predicado). Para la dimensión diacrónica los casos se etiquetaron también por décadas y semisiglos, usando la datación del corpus como base. Un resumen de la etiquetación, así como las bases de su operacionalización cuantitativa, se presentan en la tabla 1.

Como indica la línea superior de la tabla 1, se asignó un valor numérico a cada una de las diferentes alternativas combinatorias. Por ejemplo, un sustantivo en forma definida, es decir, acompañado del artículo definido, un demostrativo o un posesivo, como *la idea* en (13), recibe el valor de 0 en el criterio del determinante. De modo semejante, si el sustantivo está construido con un verbo semántico pleno, como *caber* en el ejemplo (13), recibe un valor de 0. Finalmente, si funciona como sujeto de la oración se le asigna asimismo el valor de 0. La suma de los valores de los tres criterios revela el grado de tipicidad del sustantivo en un contexto dado (valores 0 a 5). Cuanto menor la cifra, más encapsulador es el uso.

Criterio / nivel	**0**	**1**	**2**
Determinante	artículo definido demostrativo posesivo	artículo o pronombre indefinido	sin determinante
Núcleo	verbo pleno; preposición no gramatical	verbo *ser*; Preposiciones gramaticales (*a, en, por*)	verbo de apoyo (*haber, tener, hacer, dar*); preposiciones *con* y *sin*
Función sintáctica	sujeto	cd, cp, pred, desconocido	

TABLA 1. Operacionalización del estatus de encapsulador

En el último paso, los valores numéricos sumados se adscribieron a tres grados de tipicidad del uso encapsulador: los valores 0 y 1 indican uso típico; los valores 2 y 3 un uso menos típico; y los valores 4 y 5, un uso marginalmente encapsulador. Los ejmplos (13) a (18) ilustran los tres grados de tipicidad que resultan de la operacionalización cuantitativa, con un ejemplo por cada valor numérico:

Estatus de encapsulador típico

Grado 0.

(13) sólo en su cabeza trastornada puede *caber la* desatinada *idea de pedir* una gratificación estipulada para cuando llegara el caso de restituirse todos (CDH, 1801)

Grado 1.

(14) En Vinaroz *me repitieron* lo mismo; *la idea de seguir* adelante, expuesto a hallarme en Francia por fuerza y perecer allí de miseria (CDH, 1814)

Estatus de encapsulador menos típico

Grado 2.

(15) Empiezo a responder a tu última carta por donde tú la acabaste. *Confírmate en la idea de que* la naturaleza del hombre está corrompida y (CDH, 1774)

Grado 3.

(16) tiembla mi corazón y se estremece *con sóla la idea de* apartarme de Vm. cuando lo descubra. (CDH, 1785)

Estatus de encapsulador marginal

Grado 4.

(17) Simplemente María. Esa simplicidad *me daba* una vaga idea de pertenencia, *una vaga idea de que* la muchacha estaba ya en mi vida y de que, en cierto modo, me pertenecía. (CDH, 1948)

Grado 5.

(18) Subió a lo alto de las peñas, y *tuvo ideas de* precipitarse. Pero Román pensó en que Jarilla quedaba abandonada. (CDH, 1850)

Como puede observarse en los ejemplos, se nota claramente cómo *idea* en los ejemplos (13) y (14) tiene un estatus más independiente que en (17) y (18), donde el sustantivo se combina con verbos de apoyo como *dar* y *tener*, formando estructuras que se asemejan a predicados compuestos. Con esta tripartición de usos encapsuladores típicos, menos típicos y marginales, respectivamente, como base, en el apartado siguiente se emprende el análisis diacrónico de la construcción encapsuladora.

4. Análisis

En este apartado se presentarán los resultados del análisis diacrónico, empezando por algunas observaciones iniciales (§ 4.1) y siguiendo con los resultados generales (§ 4.2) que sirven para establecer una cronología para la presentación de los resultados detallados (§ 4.3). Como se verá, a lo largo del análisis, prestaré atención tanto a las diferencias (y semejanzas) entre los nueve sustantivos analizados, como a las diferencias (y semejanzas) entre las construcciones sintácticas (*de* + infinitivo, *(de) que* + cláusula, *es que* + cláusula).

4.1. Observaciones iniciales sobre los nueve sustantivos y las construcciones sintácticas

Cabe destacar, inicialmente, que todos los sustantivos no se documentan en los primeros siglos. Tampoco se insertan en las mismas construcciones sintácticas. Así, antes del XV solo aparecen casos de oraciones completivas con *condición, esperanza*, *ocasión, señal* y *sospecha*. *Causa* y *noticia*, por su parte, aparecen en los siglos XIV y XVI, respectivamente, mientras que *idea* hace su primera aparición con oraciones completivas en el siglo XVIII, y *convicción* a mediados del XIX. Entonces, queda bastante claro que para rastrear los orígenes de la construcción encapsuladora hay que investigar sustantivos diferentes de los encapsuladores típicos de la lengua actual.

Con respecto a las construcciones sintácticas, los sustantivos pueden dividirse en dos grupos: *esperanza* y *ocasión* se combinan principalmente con complementos en infinitivo, mientras que *condición*, *señal, sospecha, causa, noticia* y *convicción* prefieren completivas finitas. Solo *idea* se combina por igual con infinitivos y completivas finitas. Por su parte, *señal* se usa con una cierta frecuencia (12 %) también en la construcción con el verbo *ser* (*señal es que…*), mientras que para los demás sustantivos las cifras de esta construcción son muy bajas. Los datos iniciales del uso de cada uno de los nueve sustantivos se encuentran resumidos en la tabla 2.

Sustantivo	**Siglo de 1ª aparición**	**Construcciones sintácticas (infinitivo / finita / N ser + oración)**
condición	XIII	35 % / *59 %* / 6 %
esperanza	XIII	*63,6 %* / 35,9 % / 0,6 %
ocasión	XIII	*88 %* / 11 % / 1 %
señal	XIII	24,2 % / *63,3 %* / 12,4 %
sospecha	XIII	12 % / *87 %* / 1 %
causa	XIV	32 % / *61 %* / 7 %
noticia	XVI	11 % / *87 %* / 2 %
idea	XVIII	46 % / 46 % / 8 %
convicción	XIX	8 % / *92 %* / 0 %

TABLA 2. Panorama general de los usos de los nueve sustantivos por siglo y construcción sintáctica

A los datos generales de la tabla 2 cabe añadir algunas observaciones sobre la diacronía de los usos en las cuatro construcciones investigadas. Resulta que las preferencias de los diferentes sustantivos por una u otra cons-

trucción no son las mismas a lo largo de los siglos. Así, por ejemplo, hasta finales del siglo XIX *condición* se combina preferentemente con completivas finitas, pero a partir del s. XX es más frecuente con complementos de infinitivo. *Causa*, por su parte, aparece primero con el infinitivo (dos casos en el siglo XIV), y en el siglo XV se combina tanto con el infinitivo como con completivas finitas. Pero a partir del siglo XVI predominan claramente los casos de *causa* + oración finita. Asimismo los primeros casos de *idea* como encapsulador se dan con el complemento en infinitivo, pero a partir de inicios del siglo XIX las frecuencias de uso con infinitivo y completiva finita se igualan. Con los demás sustantivos, las preferencias generales son esencialmente las mismas en todas las épocas en las que se documentan.

4.2. Resultados generales de la operacionalización cuantitativa del uso encapsulador

La operacionalización cuantitativa de los casos extraídos del corpus permite establecer una agrupación de usos más y menos típicamente encapsuladores de los sustantivos analizados. Como quedó establecido, los seis valores numéricos posibles se han combinado para formar tres grupos de estatus encapsulador - típico, menos típico y marginal (véanse los ejemplos (13) a (18), arriba). Diacrónicamente, con un nivel de análisis por semisiglos (de cincuenta en cincuenta años), los usos típicos, menos típicos y marginales se distribuyen del modo que se presenta en la tabla 3[9].

Como revelan los colores usados para sombrear diferentes partes de la tabla 3, más abajo, la cuantificación de los usos de los seis sustantivos permite establecer cuatro periodos, que reflejan un grado cada vez mayor de usos encapsuladores típicos. Estos periodos se basan en las cifras de la tercera columna desde la derecha (en cursiva). Estas cifras corresponden a la frecuencia de los usos típicos frente a la totalidad de usos encapsuladores analizados.

1. Época medieval – 1200-1449	< 5 %
2. Siglos XV y XVI – 1450-1651	10-15 %
3. Época posclásica (siglo XVIII)	~ 20 %
4. Época moderna a actualidad – 1800-2005	> 30 %

En la tabla 3 se nota que el siglo XVII constituye un tipo de puente entre los usos antiguos, con una presencia claramente menor de usos encapsuladores típicos, y los modernos y actuales. De hecho, los datos del siglo XVII

[9] Las cifras corresponden al número total de los nueve sustantivos analizados.

Estatus de encapsulador	típico		menos típico		marginal		Suma	% de 0-1	% de 0-1	N total
	0	1	2	3	4	5				
1250	0	3	6	10	27	49	95	*3 %*	0,2 %	1756
1300	2	1	3	8	22	33	69	*4 %*	0,4 %	696
1350	0	0	3	2	7	16	28	*0 %*	0 %	343
1400	1	19	42	42	238	118	460	*4 %*	0,4 %	5165
1450	25	38	35	22	184	187	491	*13 %*	1,0 %	6467
1500	2	43	47	42	244	152	530	*8 %*	1,5 %	2944
1550	6	64	91	51	305	181	698	*10 %*	1,2 %	6057
1600	19	106	71	100	211	172	679	*18 %*	1,7 %	7424
1650	0	2	6	8	5	3	24	*8 %*	0,7 %	267
1700	11	59	32	68	50	107	327	*21 %*	2,2 %	3203
1750	37	107	92	185	88	154	663	*22 %*	2,4 %	5992
1800	49	86	52	95	57	76	415	*33 %*	3,4 %	4008
1850	98	140	90	152	87	113	680	*35 %*	3,3 %	7174
1900	92	117	99	141	74	94	617	*34 %*	3,1 %	6745
1950	171	229	179	395	223	189	1386	*29 %*	2,3 %	17 515
2000	56	39	29	80	22	22	248	*38 %*	7,4 %	1290
Suma	568	1049	877	1401	1844	1666	7404	*22 %*	2,1 %	77 046

TABLA 3. Distribución diacrónica de los diferentes estatus de los usos encapsuladores

no permiten una interpretación unívoca. En primer lugar, la segunda mitad de ese siglo arroja muy pocos casos (lo cual, por su parte, es un reflejo de la composición del CDH), por lo que las cifras no resultan del todo comparables con los de los demás semisiglos. En segundo lugar, el estatus de la primera mitad del XVII da lugar a algunas dudas. Por un lado, viendo la frecuencia de los usos encapsuladores típicos en relación a todos los usos encapsuladores, los datos de 1600-1649 se alinean con los del siglo XVIII (18 % *vs.* 21 y 22 %, respectivamente); por otro lado, si se observa la frecuencia de los usos típicos en relación con el número total de usos de los sustantivos analizados (las cifras indicadas en la columna más a la derecha de la tabla 3), la frecuencia de los usos típicos entre 1600 y 1649 se asemejan más a la de los tres semisiglos anteriores (1450-1599). Este estatus dividido del periodo de 1600-1649 entre los cuatro periodos identificados está marcado con dos colores en la línea correspondiente de la tabla 3.

Esto significa que hay que tomar la periodización con una cierta cautela. Por motivos de cronología y de presentación, en el siguiente subapar-

tado presentaré los datos correspondientes a 1600-1649 junto con los de los tres semisiglos anteriores[10] y haré caso omiso de los escasísimos casos de la segunda mitad del siglo XVII.

4.3. Observaciones detalladas

En este subapartado presentaré los datos de los sustantivos empleados en cada uno de los cuatro periodos establecidos en el subapartado anterior. Las preguntas que orientan el análisis son las siguientes: ¿Cómo son los primeros ejemplos típicos? ¿Qué sustantivos y qué construcciones se usan? ¿Hay una relación entre sustantivos, construcciones y el uso más o menos encapsulador? ¿Cómo se relacionan estos factores (los sustantivos y las construcciones) con la dimensión diacrónica? ¿Los usos típicos tienen una función más claramente textual-discursiva que los menos típicos y los marginales? La última pregunta la intentaré aclarar con un análisis particular del uso del sustantivo *señal* (§ 4.3.5). En todo este subapartado de análisis, el foco está sobre los usos encapsuladores típicos que, en ocasiones oportunas, se contrastarán con los marginales.

4.3.1. El periodo medieval (1200-1449)

En la época medieval los usos encapsuladores típicos son relativamente pocos (26 en total, de los que 20 son de la primera mitad del siglo XV; véase la tabla 3). Los sustantivos responsables de estos usos típicos son *causa* (14 casos), *condición* (3), *esperanza* (2), *ocasión* (1) y *señal* (6). Las construcciones sintácticas en que figuran los sustantivos son N *de* + infinitivo (4 casos, *esperanza* (2), *ocasión* y *señal*), N *que* + completiva (3 casos, *causa, condición, señal*), y N *ser que* + completiva (19 casos, *causa* (13), *condición* (2) y *señal* (4)). Así, la construcción que más claramente se asocia a los usos encapsuladores típicos es la que incluye el verbo *ser*. En esta construcción se insertan los sustantivos *causa, condición* y *señal*. Los ejemplos (19) a (22) ilustran los usos típicos del periodo medieval:

(19) *La* segunda *causa es que* los estrologos non cognosçen mas de mill e estrellas, e por ellas judgan (CDH, 1422)

(20) nonbres por que se nunca olujden quando leuadas fueron estas batallas & *la esperança de vençer & rreynar* & el mjedo de seer vençidos & perder rreyno. (CDH, 1284)

[10] Un motivo adicional para tratar los datos del siglo XVII junto con los de los tres semisiglos anteriores es que en ellos está aún ausente el sustantivo *idea* (que sí aparece en el siglo XVIII).

(21) Et qui fallare esta piedra, deuela mucho guardar. Et *su sennal es; que* tiene una linna por medio. Et a en ella quatro forados. (CDH, 1250)

(22) Otrosi quando el falcon a lonbrizes, *la sennal es que se le descoloran* las manos et la cera del pico (CDH, 1337)

Como revelan estos ejemplos, tanto *causa* como *señal* se emplean aquí en la construcción ecuativa N *es que*, por lo que el uso encapsulador resulta muy patente —tiene función encapsuladora tanto en el nivel semántico-cognitivo como en el discursivo-textual. Es decir, tanto *causa* como *señal* conectan diferentes elementos discursivos entre sí, según la estructura [la causa / señal de X es Y]. *Esperanza*, ejemplo (20), en cambio, no conecta con otros elementos discursivos —por ejemplo, no se especifica de quién(es) es la esperanza—, sino que el concepto de la esperanza se introduce simplemente para encapsular la noción compleja *vencer y reinar*.

Los usos encapsuladores de los ejemplos (19) a (22) contrastan claramente con los usos marginales de los mismos sustantivos que se presentan en (23) a (25). Aquí, la combinación de *esperanza, causa* y *señal* con los verbos de apoyo *haber, dar* y *fazer* da lugar a expresiones complejas que semánticamente constituyen predicados complejos:

(23) Dixo la madre del león: — *¿As esperança de estorçer* de tu grant pecado con tales palabras mintrosas? (CDH, 1251)[11]

(24) Tv eres por quien io muero e *das causa que* non muera, y eres de quien espero el galardón postrimero (CDH, 1407)

(25) e qui firiesse ell alma d»otro ol *fiziesse alguna señal que* s parasse a otra tal por ello. (CDH, 1275)

Para resumir, los ejemplos presentados permiten concluir que la construcción encapsuladora está presente ya en la época medieval, con usos bastante típicos. Los usos típicos se manifiestan especialmente en la construcción N *ser que* + oración (19 sobre 25 casos), con predominio de los sustantivos *causa* (14 casos sobre 25) y *señal* (6 sobre 25). Aunque los usos típicos son

[11] Nótese, como curiosidad, que incluso en combinación con el verbo de apoyo *haber*, en (23), parece haber una relación de identidad entre *esperanza* y la completiva de infinitivo (la *esperanza* es *estorcer*…), pese a que esta interpretación no sea la más patente por el carácter de predicado compuesto de *haber esperanza*.

poco numerosos, contrastan claramente con los usos marginales, donde los sustantivos forman predicados compuestos junto con verbos de apoyo.

4.3.2. El periodo clásico

A partir de la segunda mitad del siglo XV, los usos típicos aumentan notablemente, alcanzando un total de 298 casos (o 178, si se excluye el siglo XVII), lo cual implica un promedio en torno al 10 por ciento de todos los usos encapsuladores (véase la tabla 3). Se usan ya siete sustantivos —*causa, condición, esperanza, noticia, ocasión, sospecha* y *señal*— que se documentan en todas las construcciones analizadas: N *de* + infinitivo, N *que* + oración, N *de que* +oración y N *ser* + oración.

Si se enfocan los usos típicos, se nota que en la segunda mitad del siglo XV la construcción N *ser que* sigue siendo la más frecuente, pero desde el XVI la combinación con el infinitivo (N *de* + infinitivo) gana en frecuencia y en el siglo XVII es la construcción más frecuente. En el siglo XVI se introducen asimismo los primeros casos de la construcción N *de que* + oración. *Causa* sigue siendo el sustantivo más frecuentemente usado como encapsulador típico, con 109 sobre 298 casos (37 %); le siguen *esperanza* (68 casos, 23 %), *ocasión* (60 casos, 20 %), *señal* (28 casos, 9 %), *condición* (26 casos, 9 %), *sospecha* (5 casos, 2 %) y *noticia* (2 casos, 1 %)

Los ejemplos (26) a (32) ilustran los usos típicos de este periodo:

(26) e después sienten grande dolor, e *la causa es que* corren aí los humores corrompidos pungitivos (CDH, 1495)

(27) porque entonces reina Saturno, el cual es planeta malo, y *su condición es matar* por la destemplada frialdad. (CDH, 1585)

(28) fizo relaçion de todo lo pasado en el Andaluzia çertificandole que *la esperança de averla estava* en la yda del rey para alla. (CDH, 1481)

(29) Y siempre entre los indios *permanecía esta noticia de que* estaba allí aquel madero que ellos temían mucho. (CDH, 1605)

(30) Por mejor aviso hallo que es desterrar *la ocasión de poder pedir perdón* que pedillo y alcanzallo. (CDH, 1550)

(31) Señales. *Las señales que* la apostema esté en los riñones *es* la pesadumbre e graveza en aquella parte (CDH, 1495)

(32) No me *quedó ninguna sospecha de que* era ilusión; no vi nada, mas entendí el gran bien que hay en no hacer caso (CDH, 1562)

Con siete sustantivos diferentes, en los ejemplos del periodo clásico ya se observa cómo cada sustantivo presenta características propias. Con *causa* y *condición*, la construcción predominante de los usos típicos es con el verbo *ser*: N *ser* + oración (véanse los ejemplos (26) y (27)). En cambio, *esperanza* y *ocasión* se combinan con el infinitivo (ejemplos 28 y 30), y los sustantivos *noticia, señal* y *sospecha* con completivas finitas (29, 31 y 32). Con todos los sustantivos salvo *ocasión* se observa una relación de identidad referencial entre el sustantivo y la completiva; con *ocasión*, nombre modal, el complemento infinitivo tiene un estatus levemente diferente, pues entre uno y otro hay una relación de dependencia semántica, donde el significado abstracto de *ocasión* funciona como lo que hace posible la realización de la acción del infnitivo. Esta particularidad se aplica a la gran mayoría de los ejemplos de *ocasión*, por lo que la función encapsuladora de este sustantivo no resulta directamente comparable con la de los otros sustantivos.

Con respecto al sustantivo *señal*, se observa que, pese a que en el contexto del ejemplo (31) figura el verbo *ser*, la construcción encapsuladora no tiene el formato N *ser* + oración (véase el ejemplo (12, arriba), sino el formato N *que* + oración (véase el ejemplo 25, arriba). Esta diferencia de construcción implica que, en términos del análisis de Rodríguez Espiñeira (2018), en el ejemplo (31) la oración completiva (*que la apostema esté en los riñones*) no constituye la señal en sí, sino la hipótesis, o la creencia. La señal en (31) es *la pesadumbre e graveza…* (de ahí el plural *señales*). En cambio, en el ejemplo (22), la señal es *que se le descoloran las manos*, siendo la hipótesis o creencia que *el falcon ha lonbrizes*. En términos de Rodríguez Espiñeira (2018), los sustantivos factuales testimoniales relacionan dos entidades, una hipótesis o creencia, y una señal de esta. Como indica la comparación de los ejemplos (22) y (31), dependiendo de la construcción sintáctica, la oración completiva puede indicar o bien la señal (caso de la construcción N *ser que* del ejemplo 22), o bien la hipótesis (caso de la construcción N *que* + oración del ejemplo 31)[12].

[12] De hecho, con *causa* ocurre algo parecido, pues este sustantivo también puede poner en relación dos elementos, la causa y el efecto (o lo causado). Así, en el ejemplo 20, la oración introducida por *que* es la causa, y el efecto se deja implícito. En cambio, cuando *causa* se combina con *de* + infinitivo o completivas finitas, la completiva indica típicamente el efecto, como ocurre en un ejemplo como *saber la causa de haber así arribado* (CDH, 1589), donde el *haber arribado* es lo que ha sido causado, es decir, el efecto. Aquí, la causa queda implícita. En estos casos, no puede hablarse de relación de identidad entre el sustantivo y la completiva, pero al mismo tiempo, tampoco cabe duda de que los sustantivos

Lo que significa esto, con respecto al estatus encapsulador del sustantivo *señal*, es que la relación de identidad que se postula como criterio definitorio de la construcción encapsuladora no necesariamente se observa entre el sustantivo y la oración completiva. La «identificación» puede hacerse con respecto a otro elemento discursivo (la *pesadumbre e graveza* en (31)), de modo que la completiva puede usarse para comple(men)tar esa relación de identidad, añadiendo la información específica correspondiente a la hipótesis. Esto parece ser consecuencia de la ambigüedad de la estructura N *de que* + oración que intenta aclarar Leonetti (1993, 1999) con la distinción entre completivas apositivas y argumentales[13]. En sentido estricto, entonces, un caso como el del ejemplo (31) no debería considerarse un uso encapsulador. Pero esto sería un error, ya que tanto la encapsulación como la función discursiva de *señal* en este ejemplo es obvia. En todo caso, el hecho de diferenciar entre usos encapsuladores típicos, menos típicos y marginales hace que entre los usos típicos no aparezcan muchos casos de este tipo, mientras que abundan entre los usos marginales.

Pasando ahora a otra dimensión, la de la evolución diacrónica de la construcción encapsuladora, los datos presentados en la tabla 3 revelan una expansión de los usos típicos conforme pasa el tiempo. Sin embargo, más allá de que las cifras de la tabla 3 revelan un aumento del número de casos encontrados, hay otra forma de acercarse a la extensión de una estructura: la extensión de los contextos de uso, es decir, las colocaciones. Así, en la tabla 4 se presenta el número de diferentes núcleos sintácticos con los que se combinan los sustantivos encapsuladores en sus usos típicos.

Como revelan las cifras de la tabla 4, tanto el número de casos como el número de núcleos de los que dependen los sustantivos aumentan conforme pasan los siglos. Este aumento del número de colocaciones diferentes —es decir, el número de combinaciones de V+N distintas— puede tomarse como indicio de un aumento de la productividad de la construcción (*cfr*. Traugott & Trousdale 2013: 17-19). La diferencia más clara se observa entre los datos medievales y clásicos, si bien los números siguen aumentando hasta llegar a la actualidad. La excepción a esta tendencia de mayor productividad de los usos encapsuladores la constituyen *causa* y *señal*, que experimentan su mayor frecuencia de uso en la época clásica.

---

tienen una función discursivo-textual incuestionable. Volveré sobre la importancia de la construcción sintáctica para la función encapsuladora más adelante (§ 4.3.5).

[13] Y que, por su parte, Delbecque (2000) y Rodríguez Espiñeira (2003) refutan indicando que esa distinción no parece corresponder a la concepción que tienen los hablantes de esa estructura gramatical.

Periodo N	Medieval casos / núcleo	Clásico casos / núcleo	XVIII casos / núcleo	Actualidad casos / núcleo	Promedio casos / núcleo
causa	14 / 2 = 7	109 / 21 = 5,3	8 / 6 = 1,3	8 / 5 = 1,6	4,1
condición	3 / 2 = 1,5	26 / 7 = 3,7	7 / 4 = 1,8	23 / 11 = 2,1	2,5
convicción	–	–	–	59 / 40 = 1,5	1,5
esperanza	2 / 2 = 1	68 / 25 = 2,7	63 / 21 = 3,0	236 / 83 = 2,8	2,8
idea	–	–	14 / 13 = 1,1	337 / 156 = 2,2	2,1
noticia	–	2 / 2 = 1	32 / 21 = 1,5	142 / 54 = 2,6	2,3
ocasión	1 / 1 = 1	60 / 28 = 2,1	76 / 33 = 2,3	187 / 55 = 3,4	2,8
señal	6 / 3 = 2	28 / 5 = 5,6	10 / 5 = 2	8 / 5 = 1,6	2,9
sospecha	–	5 / 4 = 1,3	6 / 6 = 1	77 / 56 = 1,4	1,3
Promedio casos / núcleo	2,5	3,1	1,7	2,1	2,5

TABLA 4. Número de casos y de núcleos sintácticos de los nueve sustantivos a lo largo de los cuatro periodos

Cabe notar que, aparte de señalar el número de casos y de núcleos en cada periodo, también he calculado la razón entre ellos; lo que revelan las cifras resultantes es el número de casos (promedio) por cada núcleo: cuanto menor es la cifra, mayor número de núcleos diferentes. Como revelan los datos de la última línea de la tabla 4, se observa la mayor proporción de núcleos por sustantivo analizado en el siglo XVIII. Esto parece indicar que es en este siglo cuando se acaba de establecer la construcción encapsuladora tal y como la conocemos en la actualidad. Las cifras un poco más elevadas de los siglos más recientes (la columna de la Actualidad) pueden considerarse un indicio de que la construcción se ha establecido definitivamente, por lo que se observa un gran número de colocaciones fijadas, por ejemplo, puede *asaltar, confirmar(se)* y *surgir la sospecha; presentarse, proporcionar(se), perder(se), ofrecer(se), llegar, desperdiciar(se), buscar(se)* y *aprovechar(se) la ocasión*; la *noticia corre, cunde, llega, se recibe* y *se trae*; se *acaricia, concibe, rechaza, soporta* y *sugiere* una *idea*, o la *idea* puede *asustar, aterrar, atormentar* u *ocurrírsenos*, etc. En todo caso, el número de núcleos diferentes es mucho más elevado en la actualidad que en el siglo XVIII para todos los sustantivos salvo *causa* y *señal*.

Por otro lado, observando las cifras de la última columna de la tabla 4, llama la atención que los nueve sustantivos no se comporten de modo uniforme. Por ejemplo, los sustantivos que se incorporan más tarde a la construcción encapsuladora, *idea* y *convicción*, así como *sospecha*, presentan el mayor número de núcleos diferentes por caso. Inversamente, *causa* y

señal, dos sustantivos que a partir del siglo XVIII pierden importancia como sustantivos encapsuladores, presentan los valores más elevados, es decir, usos restringidos a unas pocas combinaciones altamente frecuentes (típicamente, la construcción N *ser* + oración). También *ocasión* y *esperanza* presentan valores elevados (de un 3,4 y 2,8, respectivamente); en el caso de *esperanza* esto se debe a la alta frecuencia de uso de la colocación *perder la esperanza de* (122 sobre un total de 369 casos típicos); en el caso de *ocasión* no existe una colocación tan predominante como *perder esperanza*, sino que se reparten el trabajo los verbos *presentarse, proporcionar, perder, ofrecer(se), llegar, desperdiciar, buscar* y *aprovechar*.

Como último indicio del mayor grado de independencia y productividad de los usos típicos de la construcción encapsuladora comparado con los usos encapsuladores marginales, considérese el caso de *esperanza* en (33). Este ejemplo presenta un uso marginal de este sustantivo, que tiene todas las características de un predicado complejo, la combinación *con esperanza* siendo semánticamente equivalente al verbo *esperar*:

(33) e que de alli el arçobispo se fuese a la provinçia de Toledo, *con esperança de tomar* la villa de Sepulveda que es muy fuerte (CDH, 1481)

Pues bien, en los 286 casos marginales de *esperanza* en el periodo clásico se cuentan 14 núcleos diferentes —al lado de *con* del ejemplo (33), se cuentan verbos de apoyo como *dar, haber, tener* así como la preposición *sin*— lo que implica un promedio de 20 casos por núcleo; el valor correspondiente de los usos típicamente encapsuladores de *esperanza* es de 25 núcleos por 68 casos, lo que equivale a menos de tres usos por núcleo (véase la tabla 4).

4.3.3. El periodo posclásico (siglo XVIII)

En el periodo posclásico se documentan todos los sustantivos a excepción de *convicción*. Los ocho sustantivos se insertan principalmente en dos construcciones, N *de* + infinitivo y N *de que* + oración completiva. Los casos de N *que* + oración y N *ser* + oración son escasísimos en esta época, y cuando media el verbo *ser*, la completiva es o un infinitivo o una oración causal introducida por *porque*: *la idea era apostarse…* y *la causa es porque…*

En comparación con los siglos anteriores, hay más casos y más núcleos diferentes (a excepción de *causa* y *señal*, que presentan pocos usos típicos). Como se observó arriba en relación con la tabla 4, el siglo XVIII es el periodo en el que se observa el mayor número de núcleos diferentes por sustantivo, lo cual puede verse como un indicio del pleno desarrollo de la construcción encapsuladora. Los ejemplos típicos, (34) a (41) reflejan esta situación:

(34) ¿A qué puede reducirse *la causa de que* en Portugal hubiese caído en ruina ejemplar aquel privativo comercio de la especiería, sino el orgullo y filaucía de aquella nación, a los soberbios tratan ó con los indios, que, conmovidos contra ellos, abrieron la puerta a los holandeses para su establecimiento, que se debió no tanto a la industria holandesa, cuanto a la consternación de los naturales? (CDH, 1750)

(35) Enterado el rey de estos principios, dio órdenes para que los indios no viviesen en comunidad, sino que viviesen por familias separadas, concediendo a cada uno cuatro topos de tierra, bajo *la condición de pagar* cada año ocho pesos, en dos pagos para fijar los sínodos de los curas y sueldos de ministros, gratificaciones a encomenderos, hospitales y otras pensiones. (CDH, 1770)

(36) Arizcun; hay bastantes Españoles; la lengua es infernal, y casi pierdo *las esperanzas de aprehenderla* (CDH, 1792)

(37) Llega el correo con la noticia de estar aprobada *la idea de hacer* navegable el Nalón, y que vendrán delineadores a levantar el plano topográfico (CDH, 1792)

(38) Había corrido por toda aquella comarca *la noticia de que* nuestro fray Gerundio bajaba a predicar en la función del Sacramento (CDH, 1758)

(39) También se opuso con vigor el Procurador de la Asunción reclamando los perjuicios y encomiendas, pero á pesar de todo Xerez existió el año de 1594 á costa de muchas revoluciones, y escándalos, porque Guzmán llevó adelante sus ideas aprovechando *la ocasión de que* no había gobernador general en la Provincia, que era mandado por dicho Guzmán en el Guayrá y en la Asunción por otro Teniente independiente uno de otro. (CDH, 1790)

(40) A las tres y media se nos repitió nuevamente *la señal de sondar* estando avante con el pueblo de Huacho, verificándolo con 20 brazas (CDH, 1789)

(41) Lo he leído, y me ha gustado sin duda; pero no deja de mortificarme *la sospecha de que* el sentido literal es uno y el verdadero es otro muy diferente. (CDH, 1774)

De modo parecido a lo que ocurría con los ejemplos típicos representativos del periodo clásico (ejemplos 26 a 32), también en los ejemplos del siglo XVIII se aprecian diferencias entre el tipo de encapsulación que establecen los sustantivos en (34) a (41). En primer lugar, se nota que los sustantivos *causa* y *señal*, una vez más, no parecen establecer relaciones de identidad referencial entre sustantivo y completiva, sino que se trata más de la *causa* de algo, donde el algo se corresponde con lo causado, el efecto; es decir, el efecto es

«que hubiese caído en ruina aquel comercio», pero no se especifica cuál es la causa. Pese a esto, *causa* en (34) sí parece funcionar como encapsulador de la información contenida en la oración introducida por *que* —que en Portugal ha caído en ruina el comercio de la especiería—, aunque la causa, en sí misma, parece ser el SN *el orgullo y la filaucía*. Asimismo en el ejemplo (40) con *señal*, el infinitivo no es la señal misma, sino que este se corresponde con lo señalado al «repetir la señal» [14]. Lo que indican estos datos es que los sustantivos que más usos típicos presentaban en la Edad Media ya no lo hacen en la misma medida en los últimos tres siglos. En lo que respecta a *causa* y *señal* esto cabe relacionarlo con el hecho de que la construcción sintáctica en la que más frecuentemente se encuentran estos sustantivos ya no es con el verbo *ser*, sino las de N *de* + infinitivo o N *de que* + oración.

Con los demás sustantivos, *ocasión* incluido, sí parece tratarse de relaciones de identidad referencial entre el sustantivo y la completiva (sea esta en infinitivo o de oración finita). Es decir, la *condición* es *pagar* en (35), las *esperanzas* son *aprehenderla* en (36), la *idea* es *hacer navegable el Nalón* en (37), la *noticia* es que *fray Gerundio bajaba a predicar* en (38), la *ocasión* es *que no hay gobernador* en (39) y la *sospecha* es que *el sentido literal es uno*… en (41). En suma, como indican los ejemplos presentados, en el siglo XVIII se observan usos encapsuladores bastante típicos con seis de los ocho sustantivos analizados, y esto en construcciones sintácticas donde no figura explícitamente el verbo *ser*.

Con respecto a la productividad de la construcción encapsuladora, esta puede ilustrarse con los sustantivos *esperanza* y *noticia*. En los usos típicos, *esperanza* se combina en 63 casos con 21 núcleos diferentes, frente a tan solo siete núcleos diferentes para 89 casos marginales. Las cifras correspondientes a *noticia* son 21 núcleos diferentes para 32 casos típicos frente a cuatro núcleos en 68 casos marginales. Sendos ejemplos de usos encapsuladores marginales de *esperanza* y *noticia* se presentan en (42) y (43):

(42) Borja estaba muy distante para acudir á la necesidad y no *había mucha esperanza de que* atendiese á un desorden particular de un pueblo remoto (CDH, 1786)

[14] Nótese que el sustantivo *señal* admite diferentes tipos de referentes, tanto señales materiales —una nota, una marca o distintivo, un gesto, una luz, etc.— como otro tipo de indicios inmateriales. En el ejemplo (40), donde *señal* se usa para referirse al código de señales marítimas, su probable referente es una bandera o un destello luminoso: «A las tres y media se nos repitió nuevamente la señal de sondar…», es decir se repite una señal, cuyo significado es «sondar» (lo señalado). En este sentido, la 'identificación' de la señal sería dada por el contexto extralingüístico, por el simple hecho de que los involucrados en esta situación reconocerían la señal material en cuestión.

(43) y para cortar las instancias de Eusebio pasó a *darle noticia que* el coche acababa de llegar a Londres y, según le habían referido, sin faltar cosa (CDH, 1723)

4.3.4. La época moderna (siglos XIX a XXI)

Al llegar al siglo XIX, se documentan ya los nueve sustantivos con usos encapsuladores típicos. En comparación con los siglos anteriores, y con la excepción de *causa* y *señal*, todos los sustantivos se usan con mayor frecuencia y con un mayor número de núcleos diferentes (véase la tabla 4). Asimismo la frecuencia de los usos típicos asciende ya a un promedio del 30 por ciento de todos los casos analizados (véase la tabla 3). Así, para los sustantivos *convicción, esperanza, idea, noticia* y *sospecha* los usos típicos son más frecuentes que los marginales. En cambio, para *condición, ocasión* y los ya mencionados *causa* y *señal*, los usos marginales siguen siendo más frecuentes en este periodo. Por lo que respecta a *condición* y *ocasión* esto parece deberse a que la mayoría de sus usos corresponden a unas pocas estructuras semi-fijadas, como *a/en/con condición de* (78 %) y *con, dar, perder, ser* y *tener ocasión* (55 %).

Igual que en los periodos anteriores, se documentan las cuatro construcciones investigadas, pero solo en los formatos N *de* + infinitivo y N *de que* + oración se insertan los nueve sustantivos. En el formato N *que* + oración se usan *convicción, esperanza, idea* y *noticia*, y en la construcción ecuativa N *ser* + oración se encuentran *causa, condición, esperanza, idea, noticia* y *señal*. Los ejemplos de las dos últimas construcciones son escasos: entre 1077 casos de esta época (véase la tabla 3, arriba), solo se encuentran 67 con *ser* y siete con el formato N *que* + oración. De modo parecido a lo que ocurre en el siglo XVIII, en la construcción N *ser* + oración también se documentan infinitivos con los sustantivos *condición, esperanza* e *idea*, por ejemplo, *Mi condición es saber por qué* (CDH, 1981). A diferencia de los siglos anteriores, la construcción con *ser* no es el formato más frecuente en los usos típicos de ningún sustantivo.

En los usos típicos, los sustantivos más frecuentes son *idea, esperanza, ocasión* y *noticia*, que entre ellos suman más del 80 por ciento de todos los casos (902 sobre 1077). Aunque con un número de casos más reducido, *convicción* y *sospecha* presentan su mayor frecuencia de uso en este periodo. Y, como se observó anteriormente, los usos típicos de *convicción* y *sospecha* son mucho más frecuentes que los marginales (59 y 77 *vs*. 3 y 13 casos, respectivamente).

Con *convicción, noticia* y *sospecha* predomina claramente la construcción con completivas finitas (el formato N *de que* + oración), mientras que

con *esperanza* y *ocasión* son más frecuentes las completivas de infinitivo. Con *condición* es más frecuente su uso con completivas finitas hasta finales del siglo XIX, pero en el XX predomina el infinitivo. Con *causa, condición, idea* y *señal* no se observa una preferencia ni por el infinitivo ni por las completivas finitas (N *de que* + oración). Igual que en el periodo anterior, sin embargo, en las décadas posteriores al 1800 no se observa ninguna asociación especial entre los usos típicos o alguna de las construcciones sintácticas analizadas. Aunque las construccines más frecuentes de los usos típicos son N *de* + infinitivo y N *de que* + oración, en los usos marginales de estas construcciones son mucho más frecuentes.

Finalmente, puede destacarse que en los usos típicos, los sustantivos *idea, esperanza, noticia, sospecha* se combinan con un número considerable de núcleos diferentes. Así, entre los 236 casos de *idea* aparecen 83 núcleos diferentes (siendo la preposición *ante* uno de los más destacados, véase el ejemplo (48)); para *esperanza* hay 156 núcleos diferentes por los 337 casos (*abrigar, alimentar* y *perder*, entre los más destacados); para *noticia* se documentan 54 núcleos diferentes por 142 casos (*correr, cundir, llegar, recibir* entre los predominantes); y para *sospecha* se encuentran 56 núcleos entre 77 casos (*asaltar, confirmar, surgir* y *venir* entre los más repetidos). Estas cifras pueden compararse con 11 núcleos diferentes para 137 casos marginales de *esperanza*; 3 núcleos diferentes para los 34 casos marginales de *idea*; 3 núcleos sobre 53 casos marginales con *noticia*; y cuatro núcleos diferentes para los 13 usos marginales de *sospecha*. También pueden compararse con el número de núcleos del periodo anterior (siglo XVIII), donde las cifras globales son más reducidas (véase la tabla 4).

Todo esto sugiere que la construcción encapsuladora se va expandiendo cada vez más en la lengua. Los ejemplos (44) a (52) ejemplifican los usos típicos de los nueve sustantivos de los últimos dos siglos:

(44) Ahí estaba *la causa de que* el teniente trabajara tanto allá en la finca, doblando el lomo... (CDH, 1980)

(45) del mismo modo que las leyes científicas tienen que cumplir *la condición de ser* susceptibles de comprobarse, modificarse o refutarse en un experimento posible, los principios lógicos tienen que sujetarse a la prueba de su aplicación en el proceso del conocimiento. (CDH, 2003)

(46) sintió que había comenzado a envejecer y le asaltó *la rara convicción de que* se trataba de un proceso definitivo. (CDH, 1988)

(47) Viendo el cariño entrañable que se tenian, D. Pedro concibió *la esperanza de unirlos* algun dia. Pero la muerte lo sorprendió antes... (CDH, 1847)

(48) No sintió miedo, ni nostalgia, sino una rabia intestinal ante *la idea de que* aquella muerte artificiosa no le permitiría conocer el final de tantas cosas que dejaba sin terminar. (CDH, 1967)

(49) Mis amos supieron todo cuando llegó a casa *la noticia de que* Malespina había herido mortalmente a su rival. (CDH, 1873)

(50) Oyó el marqués de Villena la palabra «marido,» y aprovechó *la ocasión de acercarse* a doña Inés. (CDH, 1850)

(51) tenía el charol de la visera tan roído y agrietado como el de la del mayor adán, y el pañuelo del bolsillo bien empapado en barro de todos los colores, *la mejor señal de que* Tolín, aunque por la categoría de su padre pudiera, y aun debiera serlo, no era de los pinturines ya mencionados, que jugaban a compás, con canicas de vidrio (CDH, 1885)

(52) Los diarios habían puesto a su alcance páginas de Lugones y del madrileño Ortega y Gasset; el estilo de esos maestros confirmó *su sospecha de que* la lengua a la que estaba predestinada es menos apta para la expresión del pensamiento o de las pasiones que para la vanidad palabrera. (CDH, 1970)

Como puede observarse, en todos los ejemplos los sustantivos parecen ejercer de encapsulador de la información expresada en la oración completiva, con la posible excepción de *ocasión* que constituye un caso límite[15]. Aunque en los ejemplos con *causa* y *señal* no hay una relación de identidad

[15] Como muy acertadamente hace notar uno de los evaluadores de este trabajo, el ejemplo (50) parece prestarse a dos lecturas: «aprovechar la ocasión *para* acercarse a doña Inés» y la encapsuladora. Nótese que en la construcción con *para* el grupo preposicional no funciona como complemento del sustantivo *ocasión* sino más bien complementa al verbo *aprovechar*. En la construcción con *de*, en cambio, hay que analizar la estructura de forma diferente, pues el complemento preposicional encabezado por *de* sí complementa al sustantivo *ocasión*. Así, se contrastan las siguientes estructuras: [aprovechar [la ocasión de acercarse a doña Inés]] y [aprovechar [la ocasión] [para acercarse a doña Inés]]. Hay que reconocer que la interpretación final de la construccion en (50) es bastante intuitiva semánticamente, teniendo en cuenta que el sujeto de la subordinada es idéntico al de la oración principal. Merece la pena contrastarlo con el siguiente ejemplo, donde el sujeto de la subordinada es diferente del de la principal: «La maestra insistió amorosamente media hora más tarde, aprovechando *la ocasión de encontrarse* el Cisne algo pensativo.» (Pardo Bazán, 1885). En este último ejemplo la relación de «identidad» resulta mucho más patente. Nótese asimismo lo que ocurre con otro verbo con el que también se combina frecuentemente *ocasión, presentarse*: «debo hablar con Martín antes que salga del escritorio de mi padre, pues en la noche puede no *presentarse la ocasión de hablarle*.» (CDH, 1862). Como no hay coincidencia de sujetos entre oración principal y subordinada, la lectura encapsuladora resulta más patente.

entre el sustantivo y la completiva, los dos sustantivos sirven para enlazar dos elementos informativos. En (44), aunque la completiva expresa el efecto, la *causa* está *ahí*, es decir, discursivamente presente. De modo parecido, en (51) la señal la constituye la oración «tenía el charol de la visera tan roído…», mientras que en la completiva se indica la hipótesis o creencia —que Tolín «no era de los pinturines»—. Es decir, aunque no haya relación de identidad, tanto *causa* como *señal* funcionan como nexos discursivos entre dos piezas de información compleja, motivo por el cual no hay razón para descalificar los ejemplos como usos encapsuladores típicos. De hecho, esta caracterización de nexo discursivo parece posible también para *ocasión* en (50), pues la ocasión que da origen a la posible acción de «acercarse» está explícitamente expresada en el contexto: es «oír la palabra 'marido'».

Por su parte, los usos de *convicción*, *esperanza, idea* y *noticia* en (46) a (49) no parecen ejercer de nexos entre diferentes elementos discursivos, sino que su función principal es la de encapsular la información contenida en las completivas. Así, se introducen los hechos de «tratarse de un proceso definitivo», «unirlos», «no permitir» y «haber herido mortalmente» como una *convicción, esperanza, idea* y *noticia*, respectivamente. Es decir, con estos sustantivos lo que parece primar es la función semántica y conceptual, no tanto la textual-discursiva, que sí está más patente en los casos de *causa* y *señal*, y asimismo de *condición* en el ejemplo (45) (*cfr*. Schmid 2000: 14).

Lo que destacan las diferencias entre los sustantivos es que la construcción encapsuladora no constituye una construcción única y homogénea. Más bien, la función de los diferentes sustantivos parece depender de su semántica léxica, de modo que los sustantivos mentales, como *esperanza, idea, convicción* tienen funciones más pronunciadamente semánticas y cognitivas, mientras que en los factuales *causa* y *señal*, y acaso el modal *ocasión*, predomina la función textual-discursiva. Así, una construcción sintáctica esquemática como son las oraciones completivas de sustantivo, sea la completiva de tipo finito o infinitivo, adquiere funciones diferentes dependiendo del tipo de sustantivo que ocupe la posición de N.

Sin embargo, hay que recordar que estas observaciones se basan en unos pocos ejemplos de un número limitado de sustantivos, por lo que no pueden considerarse como evidencia de tendencias generales. Destacan, más bien, que los usos encapsuladores pueden ser típicos en sentidos diferentes. Además, aunque sin duda es un rasgo importante, la identidad experiencial o referencial entre sustantivo y completiva no parece ser un criterio inequívoco y necesario para identificar un encapsulador nominal.

4.3.5. El caso de señal

Terminaré el apartado de análisis con unas observaciones sobre el uso del sustantivo *señal*, que tiene un papel destacado en los orígenes de la construcción encapsuladora en español. Los detalles de su uso también son relevantes para destacar la relación entre el uso encapsulador típico, el formato de la construcción sintáctica y el propio sustantivo, que acabo de mencionar al final del párrafo anterior.

Como han indicado los resultados numéricos presentados, *señal* se encuentra entre los sustantivos que más tempranamente se insertan en las estructuras gramaticales típicas de la encapsulación nominal (véase la tabla 2). Además, *señal* presenta ya en la época medieval una serie de usos claramente encapsuladores en los que tiene una función textual-discursiva prominente. Los ejemplos (53) y (54) ilustran cómo *señal* funciona como encapsulador en cuanto hace de nexo entre una observación de hecho (la señal, o prueba, en términos de Rodríguez Espiñeira 2018), y una creencia, o tesis, basada en la observación. En los ejemplos (53) y (54), se señala con cursiva el formato de la construcción sintáctica, se subraya la señal (= la observación) y se marca con negrita la hipótesis o creencia:

(53) Otrosi *la sennal* que ***el falcon ha la piedra*** *es que* non puede toller desenbargada mente et parte la tolledura (CDH, 1337, copia de la 2ª mitad del s. XIV)

(54) Et ***sil acaesçio por fanbre o por lazeria***, *la sennal es que* se le afloxan et se le acuelgan las alas et se le desparze la cola (CDH, 1337, copia de la 2ª mitad del s. XIV)

Es importante destacar que *señal* tiene la función de encapsulador anafórico de elementos discursivos en cinco de los seis casos típicos que se documentan en el periodo medieval. En el periodo clásico (1450-1649), establece una referencia anafórica en 27 de los 28 casos típicos analizados. Todo esto indica que, ya desde temprano, *señal* funciona como un encapsulador nominal destacado.

Por otra parte, como se ha visto anteriormente y según revelan los ejemplos (55) y (56), no necesariamente se establece una relación de identidad entre *señal* y las oraciones completivas que lo siguen, sean estas de infinitivo o de verbo finito. En ocasiones, las completivas indican típicamente 'lo señalado', constituyendo, por lo tanto, un caso de lo que Leonetti (1993, 1999) llama completivas argumentales. En (55) al combinarse *señal* con el verbo *poner*, se hace patente que la completiva con *de que* equivale a 'lo señalado o indicado', es decir, un argumento del predicado semántico 'señalar'. Así-

mismo, en (56), *mostrar la señal de ajustarse*, la acción expresada por el infinitivo no es la señal sino lo que se 'indica'.

De hecho, es posible ver en ejemplos como (55) y (40), arriba, un uso «cuasi deóntico» de *señal*, donde la interpretación es próxima a 'orden'. Esta lectura parece especialmente patente cuando *señal* forma parte de predicados complejos con verbos de apoyo como *dar, hacer* o *poner* y va seguida por el infinitivo. Lo que se indica o manda son casi siempre 'acciones': la señal de sondar (en el ejemplo (40)), o de disparar, combatir, virar, zarpar, etc.[16]

Así, en estos usos, aunque formalmente típicos según mi operacionalización, no se cumple el criterio de identidad experiencial entre el sustantivo y la completiva. Se señala algo, pero sin que se haga explícito cuál es la señal. Por otro lado, en los ejemplos (55) y (56) tampoco se detecta una función ana o catafórica obvia entre el complejo encapsulador (*señal* + completiva) y el contexto discursivo, lo cual supone otro indicio de que estos ejemplos no son, quizás, típicamente encapsuladores, pese a que han sido clasificados como tales por sus rasgos formales.

(55) Día 12.... El viento era muy bonancible y a veces calma; no obstante aquel Comandante se puso al pairo como a las doce y media y nosotros a su imitación, habiéndonos puesto de antemano *la señal de que* en la primera ocasión se daría fondo; en consecuencia se tomaron las dos vitaduras a las anclas de leva. (CDH, 1789)

(56) La vida con Ilona se cumplía indefectiblemente, en dos niveles o, mejor, en dos sentidos simultáneos y paralelos. Por un lado, había un estar siempre con los pies en la tierra, en una vigilancia inteligente pero nunca obsesiva de lo que nos va proponiendo cada día como solución a la rutinaria interrogante de ir viviendo. Por otra, una imaginación, una desbocada fantasía que instauraba, en forma sucesiva, espontánea y por sorpresa, escenarios, horizontes siempre orientados hacia una radical sedición contra toda norma escrita y establecida. Se trataba de una subversión permanente, orgánica y rigurosa, que nunca permitía transitar caminos trillados, sendas gratas a la mayoría de las gentes, moldes tradicionales en los que se refugian los que Ilona llamaba, sin énfasis ni soberbia, pero también sin concesiones, «los otros». ¡Ay de aquel que, a su lado, mostrara la más leve *señal de ajustarse* a sus modelos! En ese instante cortaba todo nexo, toda relación, todo nexo, toda relación, todo compromiso con quien hubiera caído en tan imperdonable debilidad y jamás volvía a ser mencionado. Iba a sumarse a «los otros», es decir, no existía. (CDH, 1988)

[16] Nótese que la gran mayoría de los ejemplos con este tipo de lectura cuasi déontica se concentran en mi corpus en un género textual específico: los relatos de viaje y militares. De hecho, casi todos los ejemplos de *poner (la) señal de* y *hacer (la) señal de* + infinitivo se encuentran en el *Diario de viaje* de Francisco Xavier de Viana (1789).

Ahora bien, aunque en (55) y (56) no haya ni relación de identidad ni una clara función discursiva, esto no significa que el sustantivo no pueda considerarse un encapsulador. Independientemente de si hay o no una relación de identidad, está claro que *señal* de algún modo subsume lo indicado en la completiva, aunque no se trate de una identificación o clasificación de esta en términos de aquella. En este sentido, en la medida en que se admite que *señal* de algún modo subsume o clasifica la información contenida en la completiva, aunque no se trate de una «identificación», podría sugerirse que se trata de un caso de coerción gramatical de parte de la construcción. Es decir, dado que la construcción N *de* + oración con muchos sustantivos, y a veces incluso con el mismo *señal*, establece una relación de identidad, o de clasificación, en la que el nombre define la oración completiva, esta interpretación puede llegar a intuirse aunque objetivamente no esté presente (*cfr*. Goldberg 1995; van Trijp. 2015).

Para terminar, y con el fin de ilustrar cómo los diferentes criterios definitorios de la encapsulación nominal distan de ser perfectos, obsérvense los ejemplos (57) y (58):

(57) Quien desprecia mi vida, *señal es que* desea mi muerte, y que la está pidiendo a voces (CDH, 1604)

(58) No estan blancas, sino rojas, *señal de que* en el camino han vertido sangre Real, pues las tiñò todas cinco. (CDH, 1600)

Ambos han sido clasificados como usos menos típicos, puesto que en ellos *señal* carece de núcleo explícito, de determinante, y se usa en estructuras formalmente fijadas: *señal es que, señal de que*. Sin embargo, lo que hacen estas expresiones fijadas es introducir oraciones subordinadas que comentan lo dicho anteriormente. Es decir, tienen una marcada función discursivo-textual. Además, aunque no hay relación de identidad entre el sustantivo y la oración introducida por *que* ni en (57) ni en (58) —en estas oraciones se indica, más bien, la hipótesis, o creencia—, la señal puede identificarse en el contexto inmediato —despreciar la vida en (57) y el color rojo en (58)—, por lo que ambas estructuras, *señal es que* y *señal de que* sí funcionan como nexos discursivos.

Los ejemplos presentados en este subapartado dedicado al sustantivo *señal*, entonces, indican, fundamentalmente, tres cosas: i) que el sustantivo *señal* tiene usos encapsuladores típicos ya en la Edad Media; ii) que la relación de identidad no es un criterio infalible a la hora de identificar usos encapsuladores; iii) que un ejemplo formalmente menos típico (o incluso atípico) no significa que un sustantivo no pueda funcionar como encapsulador.

En resumen, lo que sugieren los datos analizados en las páginas precedentes es que la encapsulación nominal expresada por medio de la construcción encapsuladora —en sus cuatro formatos— es una función altamente variable que depende de una serie de factores: i) el tipo de construcción sintáctica; ii) el contexto de uso del sustantivo encapsulador; iii) la relación entre el sustantivo y la oración completiva; y iv) la relación de la construcción encapsuladora con el contexto discursivo. Ninguno de estos factores es, en sí, suficiente para identificar, o determinar, un uso encapsulador típico de un determinado sustantivo abstracto. Para ello es necesario considerar los factores en su conjunto, lo cual solo es posible tras un análisis minucioso de su contexto de uso. Desde este punto de vista, la operacionalización formal de la construcción encapsuladora en términos de usos típicos, menos típicos y marginales supone un punto de partida útil, pero no es una solución infalible.

Hay que recordar, asimismo, que Schmid (2000, 2018) identifica tres funciones de los sustantivos encapsuladores —semántica, cognitiva y textual—. Lo que indican los datos analizados arriba es que estas funciones no se activan del mismo modo para todos los sustantivos encapsuladores. Más bien parece que algunos sustantivos son más propensos a cumplir ciertas funciones, y otros, otras. Por ejemplo, sustantivos mentales como *idea* y *esperanza* ejercen más naturalmente funciones semántico-cognitivas[17], mientras que sustantivos como *señal* o *causa* funcionan más a menudo como nexos discursivos. Dado que no se ha podido establecer una relación clara entre la construcción sintáctica y el estatus más o menos típicamente encapsulador, la conclusión parece ser que la función textual-discursiva depende del sustantivo: algunos son más aptos para establecer relaciones ana y catafóricas; otros se usan más para encapsular un contenido que se introduce en el discurso.

5. Consideraciones finales

Los datos analizados en este trabajo indican, sin lugar a dudas, que hay usos encapsuladores típicos ya en la Edad Media, y que estos usos se van haciendo cada vez más frecuentes a lo largo de los años, alcanzando su pleno desarrollo en el siglo xviii. Entonces, la respuesta a la pregunta de si puede hablarse de una construcción encapsuladora en los datos medievales, es, claramente, afirmativa. No obstante, hay que recordar que los usos típicamente encapsuladores de los sustantivos analizados son muy pocos antes del siglo xv.

[17] Recuérdese que, por tratarse de un análisis diacrónico, no mantengo la distinción entre las funciones semántica y cognitiva de Schmid (2000) sino que las trato conjuntamente.

Los primeros sustantivos que presentan usos encapsuladores son *causa, esperanza* y *señal*. Destaca particularmente el uso de los sustantivos *causa* y *señal* en la construcción N *ser* + oración. Los dos sustantivos tienen una evidente función discursiva, relacionando dos elementos discursivos entre sí (la señal y la hipótesis, respectivamente, en términos de Rodríguez Espiñeira 2018).

Con respecto a los orígenes de la construcción encapsuladora, hay un formato sintáctico que destaca sobre los demás, el que incluye el verbo *ser*: N *ser* + oración. 18 de los 25 casos clasificados como típicos en la época medieval tienen este formato; entre 1450 y 1649, la mayoría de los casos de la construcción N *ser* + oración son típicos (118 usos típicos *vs*. 63 usos menos típicos y marginales). Sin embargo, a partir del siglo XVIII la construcción con el verbo *ser* pierde importancia, quedando reducido a un porcentaje exiguo de la totalidad de casos analizados.

A partir del siglo XV, la construcción con completivas de infinitivo (N *de* + infinitivo) se hace más frecuente, y a partir de la segunda mitad del XVIII la construcción con completivas finitas (N *de que* + oración) predomina como la más frecuente de los usos típicos. Sin embargo, tanto las completivas de infinitivo como las de verbo finito siempre presentan más usos menos típicos que típicos. Más allá de la frecuencia de uso y los diferentes momentos históricos, entre las completivas finitas y de infinitivo no se observan diferencias en líneas generales. Lo que sí se observa es que algunos sustantivos prefieren las completivas finitas, mientras que otros se combinan preferentemente con el infinitivo (*cfr*. Granvik 2017a). En todo caso, esto no afecta de modo significativo el estatus más o menos típicamente encapsulador.

Ahora, en el nivel de un sustantivo particular, como es el caso de *señal*, analizado en el § 4.3.5, la construcción sintáctica sí parece importante. Las combinaciones *señal de* + infinitivo no constituyen casos típicamente encapsuladores; la combinación con *ser que*, en cambio, sí. Con *causa* ocurre algo parecido, pero con los demás sustantivos la situación es diferente.

Esto obliga a considerar la cuestión, nada sencilla, de la relación entre la construcción sintáctica, la relación de identidad, y la función (encapsuladora) del sustantivo (semántico-cognitiva y discursivo-textual). Sobre la base del análisis que he llevado a cabo en este trabajo, lo que parece ocurrir es lo siguiente: para que haya función semántico-cognitiva, es necesario que se establezca una relación de identidad entre el sustantivo y la completiva. Pero para la función discursivo-textual —para que el sustantivo funcione como nexo entre dos elementos discursivos— la relación de identidad no es necesaria. Por ejemplo, en los usos típicos de *esperanza* normalmente hay rela-

ción de identidad, pero este apenas funciona como nexo discursivo, sino que se limita a «encapsular» el contenido de la oración subordinada, caracterizándolo como una *esperanza*. En cambio, *causa* y *señal* pueden funcionar como nexos discursivos aunque entre el sustantivo y la completiva no haya relación de identidad (la identificación se da con otro elemento discursivo). Para *causa* y *señal* la construcción con el verbo *ser* es clave para su función como nexo discursivo; para *esperanza* la construcción sintáctica parece menos importante.

Aquí puede esconderse un punto que no parece haberse destacado en la bibliografía anterior que he consultado. Por ejemplo, en los trabajos que tratan la encapsulación nominal desde una perspectiva textual (*cfr*. p. ej. Borreguero 2006, 2018; López Samaniego 2011, 2018; Abad Serna 2015), la función discursivo-textual es clave, por lo que los usos que se analizan son prácticamente todos ana o catafóricos. En cambio, para Schmid (2000, 2018) la dimensión semántico-cognitiva parece más importante que la discursiva, aunque destaque las tres funciones de los sustantivos encapsuladores (2018: 112-113)[18]. Pero, como revelan mis resultados y consideraciones, los diferentes sustantivos encapsuladores tienden a ejercer diferentes funciones según su propia semántica lexica, por lo que todos no funcionan como nexos discursivos o encapsuladores conceptuales en la misma medida.

Para terminar, queda por comentar la pregunta de hasta qué punto la operacionalización formal que he propuesto permite detectar usos encapsuladores típicos. Según se desprende de la discusión de los párrafos precedentes, me atrevo a decir que el análisis formal supone un buen punto de partida para el análisis de la construcción encapsuladora —entendida en este trabajo como una entidad esquemática que abarca cuatro construcciones sintácticas, N *de* + infinitivo, N *(de) que* + oración, y N *ser* + oración—, tanto en diacronía como en sincronía. Y especialmente con base en un macrocorpus como el CDH. Los usos clasificados como típicos se corresponden, en un grado relativamente elevado, con el prototipo de encapsulación: normalmente incluyen una relación de identidad entre sustantivo y la completiva, y en ellos los sustantivos ejercen de nexo entre dos elementos discursivos. Sin embargo, debido a la ambigüedad de la construcción encapsuladora y la semántica de los sustantivos, la correspondencia no es total.

Así, teniendo en cuenta que Schmid (2000) sugiere que el estatus de sustantivo encapsulador es una categoría funcional que tiene propiedades prototípicas, terminaré planteando la siguiente escala de tipicidad de la construcción encapsuladora:

[18] Esto lo revela el subtítulo del estudio de Schmid (2000) *From corpus to cognition*.

1. Uso formalmente típico
2. Relación de identidad entre sustantivo y oración completiva
3. Función discursivo-textual (referencia ana o catafórica)

Según esta escala, los usos encapsuladores prototípicos presentarán los tres rasgos, y cuantos menos rasgos presente un uso concreto, menos típico será. He incluido la caracterización formal como primer punto en la escala dado que es el que menos problemas implica (siendo una caracterización formal), mientras que los puntos 2 y 3 incluyen una evaluación más subjetiva y más laboriosa en términos de análisis de los contextos de uso de un sustantivo dado. Además, para un análisis basado en un macrocorpus como es el del *Corpus del Nuevo diccionario histórico del español*, partir de criterios formales es una necesidad. Para una consideración global, sin embargo, sería necesario tener en cuenta los tres puntos.

Dicho esto, es obvio que el presente trabajo deja bastante que hacer para el futuro, en el sentido de que solo me he detenido en un número limitado de sustantivos, y no he analizado detalladamente la función discursivo-textual de más de uno de ellos (*señal*). Queda, pues, por describir, en profundidad, la función textual-discursiva de muchos más sustantivos; y contrastar la función textual de los usos encapsuladores típicos con los menos típicos y los marginales. Con respecto a los orígenes de la construcción encapsuladora, probablemente habría que incluir en el análisis sustantivos como *miedo, temor, razón, deseo, punto*, etc. que Schmid & Mantlik (2015) y Schmid (2018) señalan como los primeros sustantivos encapsuladores del inglés.

Corpus

CDH: Instituto de Investigación Rafael Lapesa de la Real Academia Española (2013): *Corpus del Nuevo diccionario histórico del españo*l [en linea]. http://web.frl.es/CNDHE

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

Referencias bibliográficas

Abad Serna, Silvia (2015): *Estudio contrastivo del funcionamiento semántico de los encapsuladores nominales en la prensa española y alemana: de la anáfora a la catáfora conceptual*. Tesis doctoral. Universidad Autónoma de Madrid. https://repositorio.uam.es/handle/10486/669678

Borreguero Zuloaga, Margarita (2006): «Naturaleza y función de los encapsuladores en textos informativamente densos (la noticia periodística)», *Cuadernos de Filología Italiana* 13, pp. 7395.

Borreguero Zuloaga, Margarita (2018): «Los encapsuladores anafóricos: una propuesta de clasificación», *Caplletra* 64, pp. 179-203. https://doi.org/10.7203/caplletra.64.11380

Croft, William & D. Alan Cruse (2008 [2004]): *Lingüística cognitiva*. Madrid: Akal.

Delbecque, Nicole (2000): «La estructura [el $N_{ABSTRACTO}$ *de que* + completiva]: variación formal y funcional», in G. Wotjak (ed.): *En torno al sustantivo y adjetivo en el español actual: aspectos cognitivos, semánticos, (morfo)sintácticos y léxicogenéticos*. Frankfurt am Main / Madrid: Vervuert / Iberoamericana, pp. 55-80. https://doi.org/10.31819/9783865278425-006

Goldberg, Adele (1995): *Constructions: a construction grammar account to argument structure*. Chicago: University of Chicago Press.

Gonzálvez-García, Francisco (2012): «La(s) gramática(s) de construcciones», in Iraide Ibarretxe-Antuñano & Javier Valenzuela (dirs.): *Lingüística cognitiva*. Barcelona: Anthropos, pp. 249280.

González Ruiz, Ramón (2009): «Algunas notas en torno a un mecanismo de cohesiñn textual: la anáfora conceptual», in María Azucena Penas & Rosario González Pérez (eds.): *Estudios sobre el texto: nuevos enfoques y propuestas*. Frankfurt am Main: Peter Lang, pp. 247-278.

Granvik, Anton (2012): *De* de*: estudio histórico-comparativo de los usos y la semántica de la preposición* de *en español*. Tesis doctoral. Helsinki: Société Néophilologique de Helsinki.

Granvik, Anton (2014): «Hablando de, sobre y acerca de la gramaticalización y la lexicalización: panorama diacrónico de las relaciones entre preposiciones y locuciones prepositivas dentro del campo semántico de tema/asunto», in José Luis Girón Alconchel & Daniel M. Sáez de Rivera (eds.): *Procesos de gramaticalización en la historia del español*. Madrid / Frankfurt am Main: Iberoamericana / Vervuert, pp. 77-117. https://doi.org/10.31819/9783954871988-006

Granvik, Anton (2015): «Oraciones completivas de sustantivo: un análisis contrastivo entre portugués y español», *Verba* 42, pp. 347-401. https://doi.org/10.15304/verba.42.1856

Granvik, Anton (2017a): «Oraciones completivas de sustantivo en español y portugués: ¿infinitivo u oración finita?», *Cuadernos de Lingüística de El Colegio de México* 4/1, pp. 103-180. https://doi.org/10.24201/clecm.v4i1.54

Granvik, Anton (2017b): «Accounting for syntactic variation in diachrony: the presence vs. absence of *de* before *que* in finite nominal complement clauses

in 16th and 17th century Spanish», *Belgian Journal of Linguistics*, 31/1, pp. 242-271. https://doi.org/10.1075/bjl.00010.gra

Granvik, Anton (2018): «Variación y cambio sintáctico en el establecimiento de las oraciones completivas de sustantivo en el español clásico: N *que* vs. N *de que*», in José Luis Girón Alconchel, Francisco Javier Herrero Ruiz de Loizaga & Daniel M. Sáez de Rivera (eds.): *Procesos de gramaticalización y textualización en la historia del español*. Madrid / Frankfurt am Main: Iberoamericana / Vervuert, pp. 139-171. https://doi.org/10.31819/9783954876938-006

Kabatek, Johannes (2016): «Un nuevo capítulo en la lingüística histórica iberorrománica: el trabajo crítico con los corpus. Introducción a este volumen», in Johannes Kabatek (ed.; con la colaboración de Carlota de Benito Moreno): *Lingüística de corpus y lingüística histórica iberorrománica*, Berlin: de Gruyter, pp. 117. https://doi.org/10.1515/9783110462357-001

Leonetti, Manuel (1993): «Dos tipos de completivas en sintagmas nominales», *Lingüística* 5, pp. 1-36.

Leonetti, Manuel (1999): «La subordinación sustantiva: las subordinadas enunciativas en los complementos nominales», in Ignacio Bosque & Violeta Demonte (eds.): *Gramática descriptiva de la lengua española*. Madrid: Espasa Calpe, pp. 2083-2104.

López Samaniego, Anna (2011): *La categorización de entidades del discurso en la escritura profesional: las etiquetas discursivas como mecanismo de cohesión léxica*. Tesis doctoral. Universidad de Barcelona. http://hdl.handle.net/10803/48757

López Samaniego, Anna (2018): «La encapsulación nominal en el discurso académico-científico oral y escrito: patrones de aparición», *Caplletra* 64, pp. 129-152. https://doi.org/10.7203/caplletra.64.11369

Rodríguez Molina, Javier & Álvaro Octavio de Toledo y Huerta (2018): «La imprescindible distinción entre texto y testimonio: el CORDE y los criterios de fiabilidad lingüística», *Scriptum Digital* 6, pp. 5-68.

Rodríguez Espiñeira, María José (2003): «Sobre dos tipos de completivas en frases nominales», *Verba* 30, pp. 163-202.

Rodríguez Espiñeira, María José (2015): «El sustantivo *hecho* como ejemplar de nombre encapsulador factual», in *Studium grammaticae: homenaje al profesor José A. Martínez*. Oviedo: Universidad de Oviedo, pp. 655-674.

Rodríguez Espiñeira, María José (2018): «Sustantivos con usos argumentativos testimoniales», in Ignacio Bosque, Sylvia Costa & Marisa Malcuori (eds.): *Palabras en lluvia minuciosa: veinte visitas a la gramática del español inspiradas por Ángela di Tullio*. Madrid / Frankfurt am Main: Iberoamericana / Vervuert, pp. 315-331. https://doi.org/10.31819/9783954877560-019

Schmid, Hans-Jörg (2000): *English abstract nouns as conceptual shells: from corpus to cognition*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110808704

Schmid, Hans-Jörg (2018): «Shell nouns in English: a personal roundup», *Cap-lletra* 64, pp. 109-128. https://doi.org/10.7203/caplletra.64.11368

Schmid, Hans-Jörg & Anette Mantlik (2015): «Entrenchment in historical corpora? Reconstructing dead authors' minds from their usage profiles», *Anglia* 133/4, pp. 583623. https://doi.org/10.1515/ang-2015-0056

Traugott, Elisabeth Closs & Graeme Trousdale (2013): *Constructionalization and constructional changes*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199679898.001.0001

van Trijp, Remi (2015): «Cognitive vs. generative construction grammar: the case of coercion and argument structure», *Cognitive Linguistics* 26/4, pp. 613-632. https://doi.org/10.1515/cog-2014-0074

ENTRE MIRADAS DE ASOMBRO: APORTACIONES DE LA LINGÜÍSTICA DE CORPUS AL ESTUDIO DE UNA CONSTRUCCIÓN CON LA PREPOSICIÓN *ENTRE*

The contribution of Corpus Linguistics to the analysis of a prepositional construction with entre *'amid'*

Belén López Meirama y Carmen Mellado Blanco
Universidade de Santiago de Compostela

Resumen

Los esquemas fraseológicos, por ser estructuras productivas y no cumplir con el requisito de la fijación, han sido relegados tradicionalmente a la periferia de la fraseología. En contraposición a este tratamiento marginal, los enfoques construccionales, en concreto los de corte cognitivista y basados en el uso lingüístico (*cfr*. Goldberg 2006, Bybee 2013, Butler & Gonzálvez-García 2014), presentan unos postulados interesantes para el estudio de estos patrones esquemáticos parcialmente saturados en su léxico (*partially lexically filled phrasal patterns*, Goldberg 2006). Esto permite ubicarlos, como fenómeno regular, en un punto intermedio del *contínuum* entre léxico y gramática. Además de ello, la Gramática de Construcciones, con la ayuda de las herramientas de la Lingüística de Corpus, permite la descripción integral del potencial semántico-pragmático y discursivo de estas unidades.

En este trabajo analizamos el esquema fraseológico o construcción fraseológica [*entre* + $S_{plural/corporal}$], de la cual (a) ofrecemos una descripción holística en todos sus niveles; (b) sistematizamos cualitativa y cuantitativamente las actualizaciones del *slot* $S_{plural/corporal}$ en el discurso; (c) describimos y clasificamos los colocados verbales y (d) determinamos si existe un prototipo de la construcción a partir de los datos cuantitativos y cualitativos obtenidos en el estudio de corpus, explicando cómo incide este extremo en la fijación cognitiva de la misma. Con ello hemos pretendido desarrollar una metodología de estudio de construcciones fraseológicas basada en corpus, que podrá ser implementada con otros patrones preposicionales y en otras lenguas.

Palabras clave: preposición *entre*, construcción fraseológica, Fraseología, Gramática de Construcciones, análisis basado en corpus.

Abstract

Schematic idioms, which are productive structures and therefore not fully fixed idiomatic expressions, have traditionally been relegated to the periphery of phraseology. However, constructional approaches, particularly those with a cognitivist and usage based orientation (e.g. Goldberg 2006, Bybee 2013, Butler & Gonzálvez-García 2014), contain interesting postulates for the study of these «partially lexically filled phrasal patterns» (Goldberg 2006). This approach considers such constructions to be a regular phenomenon and allows us to locate them at an intermediate point on a continuous scale between lexicon and grammar. In addition, combining the methodology of corpus linguistics with Construction Grammar allows for the description of the entire semantico-pragmatic and discourse-related potential of these units.

In the present chapter we analyse the constructional idiom consisting of the preposition *entre* 'amid' with a plural noun specifying bodily activities, which we represent as [*entre* + $N_{plural/bodily}$]. This study contains (a) a full description of all levels of this construction, (b) a systematic qualitative and quantitative analysis of the realization of the slot $N_{plural/bodily}$ in discourse and (c) a description and classification of the corresponding verbal collocates. On the basis of the results, we provide an answer to the question of whether there is a prototype for this construction and we explain how this issue is related to its entrenchment. All in all, this paper develops a general methodology for the corpus-based analysis of phraseological units which may be applied to the study of different prepositional patterns in other languages.

Keywords: preposition *entre* 'amid', constructional idiom, phraseology, Construction Grammar, corpus driven analysis.

1. Introducción: esquemas fraseológicos y construcciones fraseológicas

En la fraseología tradicional, el estudio de los patrones y la regularidad ha estado situado en un segundo plano. El centro del universo fraseológico estaba constituido por unidades irregulares desde el punto de vista formal y/o semántico, siguiendo la herencia de la corriente generativista. Por otra parte, las unidades fraseológicas con constituyentes en forma de casillas vacías, al no presentar fijación, también han sido consideradas como periféricas (vid. Ruiz Gurillo 1998: 18). Este es el caso de los llamados «esquemas fraseológicos» del tipo *de … a…* (p. ej.: *de par en par; de hito en hito*) (Zuluaga 1980: 159-160; 110-113), de los «esquemas sintácticos de formación» (García-Page 2007: 124, 2008) o de las «locuciones con casillas vacías» (Montoro del Arco

2008), como *dar* [a alguien / algo] *por* [participio] o *hacer por* [infinitivo] (p. ej.: *dar algo por perdido/sabido/…*; *hacer por venir/entender/…*).

Dado que el marco de nuestro estudio[1] es el de la Gramática de Construcciones (en adelante, CxG), preferimos el uso del término «construcciones fraseológicas» (*constructional idioms*, Taylor 2016), antes que «esquemas (sintácticos) fraseológicos», por varios motivos: (1) el término «esquema», tal y como se presenta en la fraseología española, es herencia estructuralista y se basa, fundamentalmente, en rasgos morfosintácticos, mientras que nosotras preferimos un planteamiento holístico que englobe la semántica y la sintaxis por un igual; (2) el concepto «construcción fraseológica» tiene un espectro de aplicación más amplio que los esquemas fraseológicos, que son suboracionales; (3) en la CxG tiene un papel destacado el estudio basado en corpus, de especial relevancia para analizar la productividad y la fijación cognitiva de una construcción. Seguimos en este trabajo a Taylor (2016: 11), quien define los *constructional idioms* como:

> patterns (of varying degrees of productivity and schematicity) for the formation of expressions, but whose syntactic, semantic, pragmatic, and even phonological properties cannot be derived from general principles, whether universal or language-specific.

Considerando esta definición y otros planteamientos similares, como el de Goldberg (2006: 215) sobre los *partially lexically filled phrasal patterns*, los rasgos principales de las construcciones fraseológicas los resumimos en los siguientes puntos (*cfr*. Mellado Blanco en prensa):

i. Constituyen pares indisolubles de forma y significado semántico-pragmático.
ii. Se encuentran a medio camino en el *contínuum* entre el polo léxico, formado por palabras, locuciones o refranes, y el polo gramatical, constituido por construcciones completamente esquemáticas.
iii. Ciertos constituyentes están saturados léxicamente, mientras que otros se presentan como casillas vacías o *slots* que deben llenarse en el discurso.
iv. Ni los rasgos semántico-pragmáticos ni las características sintácticas y/o prosódicas son (enteramente) deducibles de los principios gramaticales y léxicos generales de la lengua.

[1] Este trabajo se enmarca en el proyecto de investigación del MINECO *Combinaciones fraseológicas del alemán de estructura [PREP. + SUST.]: patrones sintagmáticos, descripción lexicográfica y correspondencias en español* (FFI2013-45769-P, subvencionado con fondos FEDER), el cual tiene por objetivo principal la detección y descripción de combinaciones usuales y patrones preposicionales del alemán y español.

v. Presentan gradación en la productividad del esquema según la actualización de los *slots* libres.
vi. La actualización léxica de los *slots* puede ser más o menos libre, de acuerdo con ciertas restricciones o preferencias semánticas y/o morfológicas predeterminadas.

La construcción elegida [*entre* + $S_{plural/corporal}$], con actualizaciones como *entre sollozos*, *entre risas*, *entre besos* o *entre aplausos* (así por ejemplo en *Los dos sonreímos y desayunamos con ganas*, *entre besos y miradas en silencio*), responde a esta caracterización. La hemos sometido a un exhaustivo análisis a partir de datos extraídos del CORPES XXI, con los que hemos construido un corpus de cerca de 1200 registros. El trabajo está organizado de la siguiente manera: en primer lugar mostramos la metodología seguida para identificar la construcción y para la elaboración del corpus. A continuación presentamos los rasgos generales de la construcción, para luego pasar a una descripción holística según los parámetros de la CxG. En el siguiente apartado aportamos los datos obtenidos del estudio de corpus teniendo en cuenta la frecuencia *type* y *token*, prestando especial atención a la caracterización semántica del *slot* $S_{plural/corporal}$ y a los colocados a partir de los datos proporcionados por el corpus. Todo ello nos dará las pautas para mostrar las conclusiones acerca del prototipo de la construcción.

2. Metodología y corpus

Como se ha comentado más arriba, el análisis de corpus de patrones preposicionales es uno de los objetivos del proyecto de investigación en el que se enmarca la presente investigación. Además de la detección de combinaciones usuales, una de las finalidades del proyecto es detectar posibles construcciones preposicionales a partir del estudio de frecuencia de los colocados de distintas preposiciones del español y el alemán. En muchas ocasiones, los colocados pueden agruparse en torno a criterios morfológicos y semánticos y estos grupos, a su vez, pueden estar vinculados a determinados significados de las preposiciones (*vid*. Mellado Blanco 2017; López Meirama 2017). Este procedimiento, seguido en otros estudios construccionales, como por ejemplo el dedicado a la preposición *sous* en francés (*cfr*. Lauwers 2010) o la preposición *unter* en alemán (Mellado Blanco 2018a)[2], nos ha permitido detectar

[2] En este contexto hay que citar el estudio de las preposiciones alemanas de Kiss *et al*. (2014), totalmente basado en corpus, en el que los sentidos de las distintas preposiciones se describen a partir del comportamiento textual junto a ciertos colocados sustantivos. Aunque este trabajo no persigue el estudio de construcciones, es muy útil para nosotras por el planteamiento inductivo que sigue.

en español la construcción fraseológica [*entre* + $S_{plural/corporal}$]. Para ello hemos analizado los valores semánticos de *entre* en las 500 primeras ocurrencias por frecuencia absoluta de la búsqueda <*entre* + S_{plural}> en el corpus de Mark Davies. Al intentar describir los distintos significados de la preposición, comprobamos que el valor de *entre* en ciertas combinaciones, todas ellas con rasgos comunes morfosintácticos y semánticos, se escapaba al descrito en las gramáticas consultadas (*vid*. Mellado Blanco & López Meirama 2017).

Una vez identificado el patrón preposicional y comprobado su estatus construccional, acudimos al CORPES XXI para extraer datos cuantitativos de la construcción y aportar ejemplos[3]. La recopilación de nuestro corpus culminó en julio de 2018 y se llevó a cabo del siguiente modo:

1. Búsqueda por proximidad en CORPES España: <*entre* + S_{plural}> (distancia 1, derecha del nodo *entre*) y <*entre* + A_{plural} + S_{plural}> (distancias 1 y 2, derecha del nodo *entre*).
2. Filtrado manual de cerca de 13.000 registros, en el que eliminamos:
   a. Ejemplos con sustantivos escuetos en plural pero que no se correspondían con la construcción por no presentar predicación secundaria y por no adecuarse al significado de la misma. Estas ocurrencias desestimadas presentan distintos tipos de significado, como sumativo, recíproco, local, causal o distributivo. P.e., en: *un ajuste de cuentas entre bandas rivales* (recíproco)*, La senda trepa entre prados* (local)*, Hicieron una encuesta entre personajes famosos* (local abstracto).
   b. Locuciones idiomáticas o partes integrantes de ellas, en enunciados como: *Tenía algo muy interesante entre manos*; *Entre pitos y flautas llevo varios días sin currar y por ende sin ver un chavo*; *Suele decir, entre bromas y veras, que la música es un arte inferior*. Otras locuciones excluidas son: *entre algodones*, *entre sueños*.

El número de registros que resulta después del filtrado manual alcanza aproximadamente el 10% del total de las ocurrencias que arrojan las búsquedas <*entre* + S_{plural}> y <*entre* + A_{plural} + S_{plural}>. Por otra parte, dado que uno de nuestros propósitos es estudiar la semántica de los verbos con los que coaparece el esquema, decidimos eliminar los registros sin verbo: casos en los que el esquema es modificador de un sustantivo o frase nominal, como el de (1),

[3] Puntualmente, hemos hecho algunas búsquedas en el *Corpus Diacrónico del Español* (CORDE).

y casos de verbo elidido (en didascalias, paréntesis, titulares de prensa, etc.), como el de (2), que apenas suman unas pocas decenas de registros:

(1) «No puedo más, no puedo más.» Una letanía *entre suspiros ahogados*, una voz conocida, y luego: «Preferiría morir, preferiría morir.» (M. Tena: *Tenemos que vernos*).

(2) Mari Paz.- (*Entre lágrimas.*) ¡Ay, no lo puedo remediar! (M. Rodríguez: «Las aeróbicas». *La Ratonera*).

En resumen, después del filtrado y selección de ejemplos con verbo explícito, nuestro corpus suma un total de 1169 casos (en el apartado 4.1.3 justificaremos la exclusión de los recuentos de otro pequeño grupo de registros).

3. Características de la construcción

3.1. Presentación de [*entre* + $S_{plural/corporal}$]

A continuación ofrecemos los rasgos más relevantes de la construcción según la caracterización realizada en Mellado Blanco & López Meirama (2017):

1. Tiene un doble valor: temporal —de simultaneidad— y modal, y actúa como una predicación secundaria en la que el sustantivo denota una acción paralela a la representada por el verbo modificado. Por eso en muchos contextos podría sustituirse por un gerundio: *Me lo dijo entre sollozos / sollozando* (para el concepto de «predicación secundaria», *vid. infra*).
2. «Semánticamente, los sustantivos que ocupan de manera prototípica el *slot* S se hallan en el ámbito general de la 'comunicación verbal' o de la 'expresión corporal'» (Mellado Blanco & López Meirama 2017: 255): *aplausos, gritos, lágrimas, risas, sollozos, suspiros, susurros*, etc. Formalmente, es posible la expansión del *slot*, sea a través de la subordinación (*entre gritos histéricos, entre lágrimas de emoción*) o de la coordinación (*entre vítores y aplausos, entre tartamudeos y toses nerviosas*).
3. Si bien la construcción puede coaparecer con verbos de muy diversas clases semánticas, se detecta una acusada tendencia a su combinación con verbos de comunicación y, en menor medida, de desplazamiento: *Hablaba entre risas, Se fue entre lágrimas*.
4. Pragmáticamente, tiene un claro valor enfatizador e intensificador, con una fuerte carga expresiva y emocional. Asimismo, es propia de un registro elevado y se localiza frecuentemente en textos literarios.

En las páginas que siguen, intentaremos afinar y completar esta caracterización a través de un minucioso análisis de los datos proporcionados por el CORPES XXI, con el propósito de ilustrar empíricamente lo que la CxG puede aportar al análisis de las construcciones.

3.2. Descripción holística de la construcción [*entre* + $S_{plural/corporal}$]

Para llevar a cabo una descripción integral de la construcción, es necesario abordar todos sus niveles, según plantea Croft (2001: 18) (*vid*. figura 1).

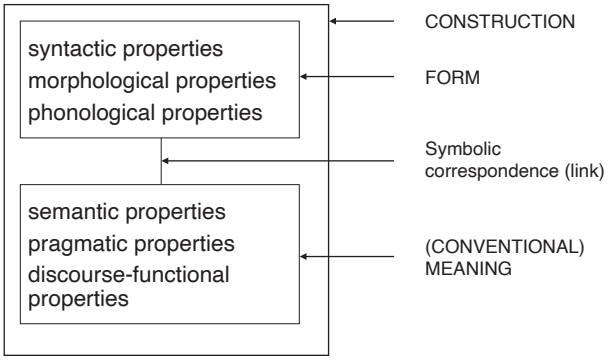


FIGURA 1. Estructura simbólica de una construcción (Croft 2001: 18)

Este esquema sirve como modelo para la descripción de todas las construcciones que conforman el constructicón de una lengua de acuerdo con parámetros que abarcan tanto su forma (propiedades sintácticas, morfológicas y fonológicas) como su significado (propiedades semánticas, pragmáticas y discursivo-funcionales). En la construcción que nos ocupa, que podemos esquematizar como se ve en la figura 2, destacamos los siguientes rasgos en cada uno de los parámetros definitorios:

[[$PREP_{entre}$] [S_{plural}]]	FORMA
[[PREDICADO] [MUY] [MIENTRAS]]	SIGNIFICADO

FIGURA 2. Representación esquemática de la construcción [*entre* + $S_{plural/corporal}$]

Características sintácticas:

i. La construcción tiene valor de predicación secundaria, siendo parafraseable por un gerundio y/o una oración subordinada temporal de simultaneidad encabezada por *mientras*. En este trabajo concebimos la predicación secundaria en sentido amplio —y no solo en relación con el adjetivo predicativo— como la que alude a las circunstancias en las que se produce la predicación primaria o principal. Puede tener significado temporal, causativo, modal, etc., suele ir en aposición y lleva a menudo adjuntos (*cfr*. Martínez García 2003: 50-55). De hecho, un tipo importante de predicados secundarios son los gerundios, con valor durativo o de simultaneidad (*cfr*. Martínez García 2003: 55; p. ej. *Saliendo de casa*, *me encontré a Juan*), que como se ha dicho más arriba pueden servir de paráfrasis de la construcción que nos ocupa. El hecho de que los ejemplos aparezcan con frecuencia entre comas en función de aposición refuerza dicho valor. Otro hecho interesante en este contexto es la existencia de argumentos dependientes de ella (p. ej. *La consoló entre susurros al oído*; *vid*. apartado 4.3), que demuestran el carácter eventivo del sustantivo deverbal.

ii. Como predicado secundario, la construcción tiene un sujeto lógico del que se predica algo. Este sujeto lógico (Agente o Experimentador), presente o implícito, puede coincidir o no con el sujeto gramatical del verbo principal del enunciado. De esta manera, establecemos dos subtipos principales de construcciones:

a. Las de sujeto coincidente: constituyen el núcleo prototípico de la construcción (subtipo 1):

(3) Los tres caminan *entre bromas* hacia la calle Hermosilla. (P. García Montalvo: *Retrato de dos hermanas*). [Los que bromean son los que caminan].

b. Las de distinto sujeto (subtipo 2):

(4) Concluye la clase *entre aplausos*. Mañana, última lección. (S. Gaviña: «¡Brava!, maestra Berganza». *ABC.es*). [Los que aplauden no concluyen].

Llama la atención que de los once sustantivos de frecuencia *token* más alta (*risas*, *lágrimas*, *sollozos*, *aplausos*, *gritos*, *bromas*, *carcajadas*, *sonrisas*, *susurros*, *gemidos*, *suspiros*) solo uno (*aplausos*) aparezca inequívocamente en ejemplos que responden al subtipo 2, de lo que se infiere que el prototipo de la construcción es el primer subtipo.

Características morfológicas:

i. Obligatoriedad del artículo cero. La aparición de determinante bloquea la interpretación de predicación secundaria y activa la de localización espacial (sea abstracta o no), como sucede en el ejemplo (9) (*vid. infra*).

ii. El sustantivo, necesariamente en plural, puede ser escueto, llevar modificación adjetiva o estar coordinado con otro u otros sustantivos (*vid.* apartado 4.2). De acuerdo con los resultados relativos a la modificación y expansión de las actualizaciones del *slot* sustantivo (*vid.* apartado 4.1), observamos una tendencia a que los sustantivos con frecuencia *token* más alta, y por ello prototípicos, sean escuetos. Por consiguiente, a menor frecuencia *token*, mayor distanciamiento del prototipo escueto y mayor tendencia a la expansión con modificadores y a la coordinación.

Características fonológicas:

Resultan de especial interés para la descripción holística de las construcciones, sobre todo las que tienen como significado constitutivo actos de habla (*cfr.* Boas 2013) y las que tienden a la realización oral (*cfr.* Schafroth 2013, Finkbeiner 2015). En el caso de la construcción [*entre* + $S_{plural/corporal}$], al no darse ninguna de las dos circunstancias, este nivel no será analizado.

Características semánticas:

El significado general de la construcción, que equivale con frecuencia a un gerundio, es el de un marco modal-temporal. El valor de marco temporal en el que se desarrolla la acción designada por el verbo principal ha surgido por una extensión metafórica del significado prototípico de la frase preposicional con *entre*, que se refiere a la ubicación de un objeto en un conjunto de cosas o entre dos extremos[4]. El análisis de los sustantivos con frecuencia *token* más elevada, mencionados más arriba, apunta hacia un significado central relacionado con estados emocionales polarizados, tanto de alegría

[4] De acuerdo con Cabezas Holgado (2013: 17), las propiedades léxicas del núcleo predicativo *entre* se resumen en dos: valor locativo y valor colectivo (p. ej. *caminar entre la gente, vivir entre ancianos*). A partir de ahí se habría desarrollado la extensión metafórica de marco temporal que presenta nuestra construcción (metáfora TIEMPO ES ESPACIO). Recuérdese, no obstante, que tanto el valor local abstracto como el valor temporal de simultaneidad de *entre* (*cfr. entre semana*) ya existían en su predecesor latino *inter* (*cfr. inter graecos*: 'entre los griegos', *inter noctem*: 'durante la noche'; *cfr.* Hernández Díaz 2014: 1644 y 1652). Por las calas realizadas en el CORDE se constata que la construcción ha pasado de una localización imprecisa a una conceptual. Junto con otros sustantivos similares, las risas y las lágrimas se han usado metafóricamente para describir de manera plástica situaciones que enmarcan un acontecimiento. De ahí probablemente se pasó a una predicación secundaria: describir una circunstancia desembocó en describir una acción simultánea.

como de tristeza (campos semánticos *reír*, *bromear*, *llorar*) y actos sociales (*aplausos*). Y más en detalle, el significado concreto en cada realización depende de las características semánticas del sustantivo que actualice el *slot*.

La morfología deverbal prototípica de los sustantivos de la construcción y su valor eventivo propician que esta pueda ser interpretada con un significado verbal de simultaneidad, junto a su valor modal. Así, por ejemplo, la secuencia *entre risas* en (5) podría equivaler al gerundio *riendo* o a la cláusula temporal *mientras reían*. Lo mismo sucede en los casos de coordinación, como en (6).

(5) Juan le pidió que se quitase la gabardina y, al descubrir el uniforme, todos aplaudieron *entre risas*. (O. Aibar: *Los comedores de tiza*). [*Cfr.*: aplaudieron riendo/mientras reían].

(6) *Entre lloros e insultos* empujaba a todos aquellos que se ponían en su camino. (A. Torres Blandina: *Niños rociando gato con gasolina*). [*Cfr.*: los empujaba llorando e insultándolos].

En los casos de expansión del sustantivo, el adjetivo posnominal puede interpretarse con valor adverbial:

(7) Automáticamente, el grandullón protestante se desplomó *entre gritos histéricos*, y todo el patio iba pregonando: *«Le petit espagnol a mordu Lanson!»*. (A. Boadella: *Memorias de un bufón*). [*Cfr.*: se desplomó gritando histéricamente].

Características pragmáticas:

La construcción presenta un elevado grado de expresividad que contribuye a intensificar el significado del enunciado. Esta expresividad se ve respaldada (1) por el propio significado emocional de las actualizaciones léxicas del *slot* sustantivo, (2) por los modificadores adjetivos, a menudo intensificadores, (3) por la coordinación, puesto que los elementos coordinados son a menudo del mismo campo semántico y de significado similar o antitético, lo cual refuerza icónicamente la expresividad inherente a la construcción en la que se insertan. Con la construcción, el hablante persigue resaltar expresivamente la circunstancia emotiva que acompaña al sujeto, a menudo mientras emite un mensaje o, en menor medida, al desplazarse (*vid.* apartado 4.3).

Características discursivo-funcionales:

La construcción aparece preferentemente en el registro literario. En consonancia con este nivel, en las instancias se observan numerosos juegos crea-

tivos con el lenguaje, muchos de los cuales afectan a las actualizaciones del *slot* sustantivo, de ahí su diversidad (297 sustantivos distintos). Ello también explica la abundancia de largas modificaciones adjetivas (p. ej. *entre procaces comentarios*, *deseos rutinarios y algo viscosos*, *verbalizados con la mayor naturalidad*) y las coordinaciones incongruentes desde un punto de vista semántico y funcional (p. ej., *trapichear entre traspiés y adulación*).

4. Análisis de la construcción [*entre* + $S_{plural/corporal}$] basado en corpus

4.1. Análisis del *slot* sustantivo $S_{plural/corporal}$

4.1.1. Forma del slot: *la expansión del sustantivo*

Una de las características tradicionalmente asociadas a los fraseologismos es la fijación formal. Al respecto, [*entre* + $S_{plural/corporal}$] manifiesta dos estrictas limitaciones formales en el *slot* sustantivo: el número plural y la ausencia de determinación. En lo que concierne a la primera, los sustantivos susceptibles de ocupar el *slot* no pueden combinarse en singular con la preposición *entre* a no ser que estén coordinados, en cuyo caso estamos ante una construcción diferente, como (8), donde *entre* aporta el valor de 'estado intermedio'. Por su parte, como ya hemos señalado, la presencia de un determinante entre la preposición y el sustantivo bloquea el valor de predicación secundaria que caracteriza la construcción aquí analizada: en (9), el valor asociado a *entre* es el de 'localización':

(8) A Elsie se la ve con un gesto *entre sonrisa y mueca*, lo que promueve un efecto agrio. (M. Mactas: *Una mujer peligrosa. En un lugar envenenado de la Tierra*)

(9) Voces de Estate quieto se mezclan con risas infantiles mientras Pepita camina más despacio de lo que desea debido al peso de la niña; y *entre las risas*, descubre al hijo pequeño de Reme. (D. Chacón: *La voz dormida*) [Descubre al hijo de Reme entre los niños que ríen].

Los casos de expansión del sustantivo, sin embargo, no son infrecuentes. Los recuentos que hemos llevado a cabo en CORPES XXI han aportado los siguientes datos: de los 1169 registros analizados, 680 son casos de sustantivo escueto; 240, de sustantivo modificado y 249, de construcción coordinada, lo cual significa que los casos de expansión superan el 40% de la muestra (alcanzan el 41,83%):

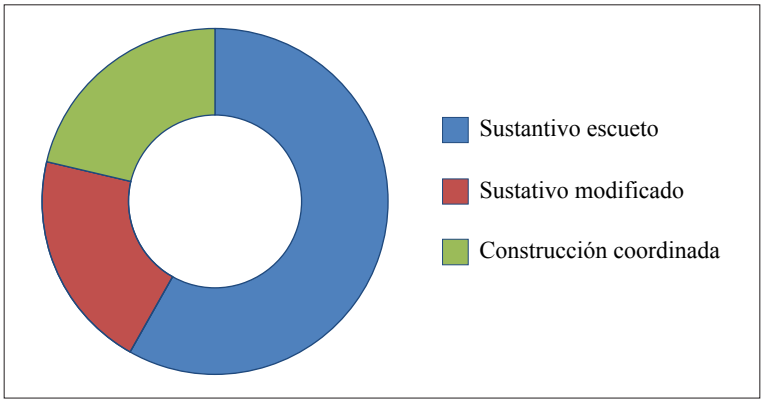


GRÁFICO 1. Tipología formal de los sustantivos

Respecto a la modificación del sustantivo, a menudo sirve para la especificación de su denotación, sea a través de un adjetivo o frase adjetiva: *entre balidos lastimeros*, *entre cánticos religiosos*, *entre comentarios aduladores*, *entre gritos completamente indecentes*, sea a través de una frase preposicional: *entre espasmos de dolor*, *entre gemidos de amor*, *entre lágrimas de emoción*. Por tener un significado muy general, algunos sustantivos solo se emplean en la construcción si están modificados: *entre frases ininteligibles*, *entre gestos de escepticismo*, *entre muestras de júbilo*, *entre palabras de despedida*.

La modificación del sustantivo puede producirse también por un adjetivo antepuesto. De la búsqueda en el corpus de la secuencia <*entre* + ADJ_{plural} + S_{plural}> resultaron 812 registros, de los cuales 69 se ajustan por su significado al de la construcción. El análisis de la expansión interna de la construcción resulta especialmente interesante desde el punto de vista de la fijación para ver el grado de cohesión entre la preposición y el sustantivo, que se supone alto porque estamos ante sustantivos que no admiten determinante (*cfr*. Mellado Blanco 2018a). Los resultados apuntan a que un 28,7% de los casos de modificación están representados por adjetivo antepuesto, lo cual indica que el grado de cohesión entre preposición y sustantivo no es absoluto. Algunos adjetivos antepuestos poseen valor de epíteto, es decir, no aportan prácticamente ningún nuevo rasgo al sustantivo. Más bien al contrario, redundan en algún sema inherente a su significado (p. ej. *confusos murmullos*, *sonoras carcajadas*, *oscuros remordimientos*, *confidentes susurros*). Estamos mayoritariamente ante cuantificadores intensificadores que acentúan la expresividad propia de la construcción, del tipo *grande*, *enorme*, *fuerte*, *numeroso*,

firme, *incontenible* y a veces podríamos hablar de colocaciones, en las que el colocado adjetivo aporta la función MAGN al sustantivo: *incontenibles risas*, *desgarrados lamentos*. Otros adjetivos tienen un matiz subjetivo negativo: *aburridas explicaciones*, *asquerosos estertores*, *torpes palabras*, o bien positivo: *afectuosas risas*, *admirativos aplausos*. Se observan, igualmente, algunas proyecciones metafóricas, como en *ebrios forcejeos* y algunos evidenciales epistémicos (*evidentes muestras de nerviosismo*).

En términos generales podemos afirmar que el registro literario característico de la construcción propicia el uso ornamental del adjetivo como epíteto. En este contexto, conviene recordar que, en palabras de Maldonado (2010: 74), «[l]os casos de dislocación del adjetivo en que antecede al nombre en vez de sucederlo, corresponden consistentemente a fenómenos de subjetivización». Y precisamente son la subjetivización, la fuerte emocionalidad y la expresividad los rasgos que definen pragmáticamente nuestra construcción.

En cuanto al tipo de sustantivos proclives a una modificación por adjetivo antepuesto, resaltan los que denotan algún tipo de actividad o efecto sonoros, que se clasifican de la siguiente manera[5]:

1. Expresión acústica, ya sea de discurso articulado o no articulado, relacionada con actos sociales: *aplausos*, *bravos*, *aclamaciones*, *agradecimientos*, *exclamaciones de aprobación*, *murmullos*. En algunos casos con tinte negativo: *protestas*, *himnos de tristeza*, *críticas*.
2. Discurso: *declaraciones*, *conversaciones*, *palabras*, *acusaciones*, *explicaciones*, *descalificaciones*, *sugerencias*, *indirectas*, *comentarios*, *rumores*, *susurros*.
3. Expresión acústica relacionada con emisiones de sonidos resultado de estados emocionales negativos: *quejidos*, *estertores*, *quejas*, *lamentos*.

En relación con este último grupo detectamos sustantivos del campo del dolor, como *dolores*, *padecimientos*, *sufrimientos*, *convulsiones*, *(dolorosos) esfuerzos*, *escenas de dolor*, *muestras de dolor*. En esta descripción destacamos, por tanto, sustantivos cargados de una fuerte emocionalidad en sí mismos, muchos con un sema acústico y/o de dolor, por un parte, y por otra parte relacionados con el discurso articulado. Estas observaciones avalan nuestra

[5] Llama la atención aquí que el sustantivo con más frecuencia *token*, *risas*, apenas aparezca con modificador adjetivo antepuesto: la búsqueda <*entre* + X + *risas*> arroja una mayoría de actualizaciones del *slot* X como determinante, pero apenas como adjetivo (algunas excepciones son *entre apagadas risas, entre eufóricas risas, entre irónicas risas*). Este dato concuerda con el elevado grado de fijación y cohesión de la unidad *entre risas*, avalado por su alta frecuencia *token*.

tesis principal de que la construcción analizada sirve fundamentalmente para expresar con intensificación los estados anímicos y los puntos de vista (de aprobación o desaprobación) que acompañan la realización de las actividades designadas por el verbo.

Siguiendo con los casos de modificación hallados en el corpus, constatamos que en algunas ocasiones esta se realiza mediante una frase preposicional que acompaña al sustantivo y denota al agente de la acción asociada a este, como puede apreciarse en los ejemplos siguientes:

(10) Siguen llora que te llora, llenando pañuelos de lágrimas reales que caen al suelo *entre risas de los demás niños que ven el ensayo*. (J.L. Alonso de Santos: *¡Viva el teatro!*)

(11) Tras su puesta en libertad eran recibidos en los pasillos del Juzgado *entre aplausos de los suyos*. (S. Pena: «Operación Campeón. Seguirán imputados sin medidas cautelares». *El Mundo.es*)

Si bien con menor frecuencia, también hemos detectado casos en los que el sustantivo se acompaña de otros argumentos:

(12) [Se preguntaba] por su empeño en invitarle a sentarse con él en una de las terrazas para contemplar a la gente, *entre comentarios algo embarazosos a uno de los camareros*. (R. Bodegas: *El ciclista solitario*).

(13) Un recluso español saca un «pincho carcelario» hecho con un clavo y ataca a uno de los internos marroquíes que había resultado herido cuatro días antes. «Me había faltado al respeto», se justificó *entre insultos al «moro»* ante los funcionarios que lo detuvieron. («Seis presos españoles agreden a tres marroquíes en una cárcel». *Público.es*).

Estas modificaciones refuerzan la interpretación de la construcción [*entre* + $S_{plural/corporal}$] como una predicación secundaria.

Respecto a la expansión por coordinación, en (Mellado Blanco & López Meirama 2017) destacamos la presencia en los corpus no solo de binomios (*entre bromas y risas*, *entre caricias y besos*, *entre gemidos y dolores*, *entre lágrimas y suspiros*, *entre palmas y vítores*), sino también de trinomios (*entre hipidos*, *mocos y temblores; entre quejas*, *lamentos y exabruptos; entre risas*, *espasmos y bocanadas*), haciendo notar que «la acumulación de sustantivos refuerza el valor expresivo de la construcción y contribuye a lograr un clímax emocional» (Mellado Blanco & López Meirama 2017: 258). En ocasiones, los elementos coordinados tienen significados similares (*entre*

alabanzas y halagos, *entre balbuceos y vacilaciones*, *entre gritos y voces*, *entre risas y sonrisas*, *entre toses y carraspeos*), en una redundancia acumulativa e icónica; o bien antitéticos (*entre sollozos y sonrisas*, *entre empujones y bromas*, *entre injurias y aleluyas*), lo cual aporta un valor de polarización entre dos extremos (*ibídem* 2017: 258). Asimismo, en CORPES XXI se registra algún caso de reiteración: *entre aplausos y más aplausos* o de inclusión de un focalizador: *entre saltos*, *volteretas y hasta canciones*.

La coordinación, por otra parte, permite enriquecer la construcción a través de algún valor añadido, como el distributivo (*entre sonrisas de ella y encogimientos de hombros de él; entre oles femeninos y risas masculinas*) o el incremental (*entre buenas palabras y mejores deseos; entre insultos y más golpes del vecindario*), o contribuye a la expresividad a través de algún recurso retórico, como el quiasmo (*entre tibios aplausos y risas irónicas*). En otras ocasiones, el segundo elemento del binomio no es un sustantivo, sino un enunciado completo sustantivado en estilo directo, como en *entre risas y un «ya te llamaremos»*.

Este último fenómeno está en estrecha relación con el concepto de *coerción*, al que por su relevancia vamos a dedicar unas líneas.

La *coerción*, término de tintes generativistas impulsado por Pustejovsky (1991), ocupa un lugar relevante en la CxG (*cfr*. Goldberg 1995, Michaelis 2003a, 2003b y 2011, Boas 2011, Gonzálvez-García 2013), especialmente en su corriente cognitiva[6]. La esencia de la coerción se basa en un desajuste entre un término y la construcción donde se integra, de tal manera que, siguiendo el «principio de anulación» (*override principle*) formulado por Michaelis (2004), la semántica de la construcción se impone sobre la semántica del término insertado en ella. Este fenómeno afectaría en principio a todas las construcciones parcialmente esquemáticas, como recuerda Michaelis (2003b: 176): «Coercion is instead a natural by-product of type selection. Any construction that selects for a specific lexical class or phrasal daughter is a potential coercion trigger».

Por otra parte, la coerción es un fuerte argumento a favor de la indisolubilidad entre significado y estructura, como rasgo que propugna la CxG para todas las construcciones. En este sentido, Lauwers & Willems (2011: 1220) apuntan que desde el punto de vista de la CxG:

[6] En la Lingüística Cognitiva, el término *coerción* se refiere al ajuste reinterpretativo que se da en caso de conflicto «between the semantic properties of a selector (be it a construction, a word class, a temporal or aspectual marker) and the inherent semantic properties of a selected element, the latter no being expected in that particular context». (Lauwers & Willems 2011: 49).

> Coercion constitutes a major argument in favor of the existence of constructions as independent form/meaning pairings, since it can be used as a heuristic means to discover the independent constructional semantics. If a construction is able to change the meaning of a lexical item that occurs in it, then one is entitled to say that the construction has a particular meaning on its own, irrespective of the lexical items that instantiate the construction.

El tipo de coerción que mayoritariamente aparece en nuestra construcción es el «exocéntrico», siguiendo la clasificación de Michaelis (2004: 7), dado que el elemento que la ejerce no es el núcleo de la construcción, sino el primer sustantivo coordinado, cuya semántica construccional junto a la preposición arrolla e impregna el significado del sustantivo «coercionado», obligando así a su reinterpretación. Este es el motivo por el cual el elemento coercionado aparece siempre en segundo, tercero o cuarto lugar, y nunca en el primero. Cuando el elemento discordante con la construcción aparece en primer lugar del binomio (*Echarse encima de alguien entre manos y dentelladas*) no se da la coerción, sino que se activa el significado local de *entre* (*entre manos*: 'en medio de muchas manos'), que coexiste con el de la construcción (*entre dentelladas*: 'dando dentelladas').

Teniendo en cuenta, entonces, el principio de coerción, observamos que en ciertos contextos se yuxtaponen sustantivos heterogéneos que pasan a ser interpretables con valor temporal-modal, y no como meros complementos circunstanciales locales:

(14) Pero mientras quitamos el tapón del desagüe *entre gritos, consignas y pancartas* (da gusto así, contra el PP), resulta que la cartelera de estrenos se convierte en una fiesta. (O. Marchante: «Batman se disfraza de Bruce Wayne». *Una de piratas*. *www.abcblogs.abc.es/cine*)

El sustantivo *pancartas*, al aparecer tras *gritos* y *consignas*, pasa a designar de manera implícita la acción de portar pancartas en una manifestación: la cartelera se llena de estrenos mientras los manifestantes gritan, lanzan consignas y agitan pancartas. De faltar los otros sustantivos, *entre pancartas* debería interpretarse en sentido meramente local de ubicación dentro de un conjunto, como sucede en (16):

(15) *Entre pancartas y vehículos con publicidad variada, además de grupos fuera de concurso*, desfilaron 44 agrupaciones que, a partir del próximo lunes, competirán en el concurso oficial del Teatro de Verano. («Volvió Momo y sus colorines». *El País*. CORPES XXI, Uruguay)

Otras veces, la coordinación permite la inclusión de sustantivos en singular con la misma interpretación: *[Corrían por las calles de Roma] entre risas y desenfreno* ('riendo y desenfrenándose, desmandándose'); *[Entraría en la fiesta] entre vítores*, *aplausos y lluvia de confeti* ('mientras los demás vitoreaban, aplaudían y tiraban confeti'); *[Respondió a las preguntas] entre lágrimas y voz baja* ('llorando y hablando en voz baja'). Estos sustantivos en singular, claro, deben ser compatibles con la construcción; los detectados en el corpus suelen ser continuos (*lluvia*, *tabaco*) o eventivos (*desenfreno*, *coqueteo*), lo cual favorece su interpretación como predicaciones.

En alguna ocasión, incluso, el segundo o tercer miembro de la coordinación es una frase nominal con determinante: *entre gritos*, *carreras*, *y alguna otra reprimenda de las educadoras; entre risas y un poco de coqueteo*. En estos casos se hace evidente que nos alejamos del núcleo de construcción. En (16) destaca, además, el valor sumativo de *entre*:

(16) La actriz va llevando al público de forma sutil al terreno de los grandes textos, *entre sonrisas, alguna que otra canción y muchas carcajadas*. («Amores menudos». *Artez*.)

4.1.2. Expansión del slot sustantivo y redes entre construcciones

En relación con lo dicho en el punto anterior, debe tenerse en cuenta que la coordinación propicia la presencia de otros valores de *entre*, que pueden entremezclarse con los de la construcción objeto de nuestro análisis, hasta el punto de que en determinados contextos se superponen. Desde el punto de vista de la CxG, conviene señalar que la propia idiosincrasia de las construcciones favorece la multiplicidad de sentidos por la superposición que puede darse con construcciones cercanas. Así, basándonos en la apreciación de Goldberg: (1995: 31): «Constructions are typically associated with a family of closely related senses rather than a single, fixed abstract sense. Given the fact that no strict division between syntax and the lexicon is assumed, this polysemy is expected [...]», descubrimos que la construcción analizada puede presentar en ciertos contextos un valor causativo, como en (17), influida probablemente por el valor causativo extraconstruccional de *entre* (p. ej. en *Entre el frío y la falta de calefacción, vas a caer enfermo*), o un valor locativo, como en (18):

(17) [...], el acordeonista y el violinista podrían haber sido nuestro vecino del quinto, porque no se les distinguía una triste nota. *Entre brincos y desenfreno*, la acústica puede parecer un detalle menor, pero esta pandilla de emigrantes hizo bien en retirarse tras poco más de una hora. (F. Neira: «Gogol Bordello o el desmadre». *El País.com*)

(18) La estampa del familiar emparedado persigue al adolescente [...], cuando en las tinieblas se le aparezca el rostro lívido que hace días descansaba en el féretro sobre el colchón de la cama más grande de su casa, *entre cirios y susurros*[7]. (M. Longares: «Personajes. Los difuntos». *La ciudad sentida*)

En otras ocasiones, como hemos apuntado más arriba, se vislumbra un valor sumativo (19) o bien de alternancia (20), ambos presentes en la preposición *entre* fuera de la construcción (p. ej. *Entre parientes y amigos suman 23; Hacen turnos entre el padre y la madre*):

(19) Por su parte, Chris Bosh, recibido *entre aplausos y abucheos a partes iguales*, culminó otra inspirada noche del 'Big Three'. («Derrota y conato de crisis en LA». *El Mundo.es*)

(20) La noche, semejante a las anteriores, transcurrió *entre bostezos, juegos de parchís y oca*, (no quisieron que les instalasen la televisión para disfrutar del ambiente, y ahora se arrepentían de ello). (P. Paz Pasamar: «Historias Bélicas. La casa rural». *Historias Bélicas.)*

Las relaciones que mantienen las construcciones entre sí han sido estudiadas desde distintos ángulos de la CxG. En primer lugar, se habla de relaciones verticales de herencia en el *contínuum* léxico-gramatical (*cfr*. Goldberg 1995, 2006), de acuerdo con las cuales las construcciones menos esquemáticas, es decir, las más saturadas léxicamente, heredarían las propiedades y rasgos funcionales de las más esquemáticas, que son las más próximas al polo gramatical. En este contexto, podríamos concebir el valor causativo, sumativo y el de alternancia como «herencia» del uso de *entre* en combinación libre con sustantivos, ya estén en singular o plural. Otro caso de relación de herencia y relación vertical lo encontramos con el valor espacial 'en medio de un conjunto de cosas', ya sea estático o direccional, que sería el valor prototípico inicial de la construcción (*vid*. nota 4). Así, en (21), el sustantivo *cascotes* no denota metonímicamente ninguna actividad (como sucedería en el caso de que se diera coerción), sino tan solo alude a un espacio físico.

[7] A nuestro juicio, la posición de los sustantivos influye en la interpretación de secuencias como esta, en la que se activa el valor locativo extraconstruccional junto al construccional, frente a los casos de coerción, vistos más arriba en el § 4.1.1., pues para que haya coerción consideramos necesario que el elemento que no encaja semánticamente en la construcción se sitúe al final de la misma. En otras palabras, la anteposición de elemento que no encaja con la construcción bloquea la coerción, mientras que su situación final no la asegura (*vid*. ejemplo 21).

(21) La cabeza le arde. El mundo arde consumido a su alrededor. «Aquí estoy a salvo de todo», y camina y tropieza *entre espasmos y cascotes*, «excepto de mí mismo». (J. Avilés: *Constatación brutal del presente*) (*Cfr*.: tropieza convulsionándose / tropieza entre los cascotes).

En segundo lugar, al margen de las relaciones verticales, algunos autores como Van de Velde (2014: 147), Gonzálvez-García (2014 y 2017) o Traugott (2018: 19-20) constatan relaciones horizontales interconstruccionales, en virtud de las cuales, «a particular construction may be partly motivated in relation to its neighbours» (Van de Velde 2014: 147). Tal sería el caso de la relación que entabla [*entre* + $S_{plural/corporal}$] con [*entre* S1 *y* S1] en su valor temporal, tal y como ha sido estudiada por López Meirama (2017). Esta construcción, que cuenta con dos valores básicos, uno local y otro temporal, presenta en la dimensión temporal un uso que se acerca mucho a la de nuestro objeto de estudio. Según el DEA (2011², s.v. *entre*), la preposición *entre* «precediendo a dos nombres iguales unidos por Y o a un nombre en pl, expresa circunstancia reiterada o continuada que acompaña a la acción del verbo», significado que ilustra con ejemplos como los siguientes: *El forastero respondió, entre vaso y vaso, que venía de muy lejos; Recita las máximas de Mao entre whisky y whisky*. Estamos aquí de nuevo ante una predicación secundaria en la que los sustantivos actualizadores del *slot* S1 evocan acciones repetidas realizadas simultáneamente o de manera intercalada con respecto a la acción designada por el verbo principal.

Cuando el *slot* sustantivo de [*entre* S1 *y* S1] se actualiza con nombres relacionados con los campos semánticos 'comer', 'beber (casi siempre alcohol)' y 'fumar' (p. ej., *bocado*, *sorbo*, *calada*) detectamos puntos de interconexión semántica con [*entre* + $S_{plural/corporal}$]. En ambas construcciones, los sustantivos escuetos, aunque sean concretos y designen cosas, evocan predicaciones por adaptación metonímica al contexto. A través de (22) y (23) podemos constatar la cercanía de significados, sobre todo cuando [*entre* + $S_{plural/corporal}$] presenta sustantivos coordinados (*entre jarras de vino y cazuelas de sopas*, *entre vinos y tapas*), lo cual, además, sucede en el 68% de los registros de este grupo, cifra que contrasta con el 21% global (*vid*. gráfico 1):

(22) Fue allí, *entre tazas de té y gin-tonics,* donde me contó la historia de Miralles. (J. Cercas: *Soldados de Salamina*)

(23) Se acercaba a nosotros la gente, comentaba *entre copa y copa, entre café y ginebra*. (A. Gala: «LIRIA», *Los invitados al jardín*).

Vemos aquí como el significado de simultaneidad de [*entre* + $S_{plural/corporal}$] confluye con el de intercalación, propio de [*entre* S1 *y* S1]. Asimismo, esta segunda construcción se impregna del valor temporal de simultaneidad de [*entre* + $S_{plural/corporal}$].

Por último, la interconexión de [*entre* + $S_{plural/corporal}$], especialmente en su versión coordinada, y [*entre* S1 *y* S1], las dos con valor temporal e intensificador y como predicaciones secundarias, favorece secuencias como la de (23), *entre café y ginebra*, en las que *entre* introduce una estructura coordinada de dos sustantivos escuetos distintos en singular (*entre* + $S1_{singular}$ *y* $S2_{singular}$). Esta estructura puede considerarse un híbrido formal de las dos construcciones comentadas, con un valor similar: significado iterativo de simultaneidad de una acción de 'comer', 'beber' o 'fumar' (a la que se alude metonímicamente mediante $S1_{singular}$ *y* $S2_{singular}$), en la que se intercala la acción descrita por el verbo principal.

4.1.3. Semántica del slot $S_{plural/corporal}$*: tipología del sustantivo*

Según se expuso ya en Mellado Blanco & López Meirama (2017), los sustantivos, si bien no son sustantivos eventivos propiamente dichos, sí designan lo que la lexicografía tradicional define como «acción (y efecto) de». Así, muchos se sitúan en el ámbito del intercambio lingüístico, sea por su contenido (*insultos*, *acusaciones*, *exclamaciones*) o por su manifestación física (*gritos*, *susurros*, *murmullos*). Otros denotan determinados movimientos o emisiones corporales cuya realización implica ruido (*hipos*, *jadeos*, *ronquidos*) y están preferentemente asociados al llanto o a la risa (*carcajadas*, *sollozos*). Finalmente, algunos pueden denotar manifestaciones gestuales de diverso tipo (*aplausos*, *besos*, *miradas*). Esta caracterización, relacionada con actividades realizadas con el cuerpo humano (o animal) en sentido general, justifica la elección del adjetivo *corporal* en la formulación de la construcción [*entre* + $S_{plural/corporal}$].

Siguiendo con la descripción del *slot* sustantivo, en Mellado Blanco & López Meirama (2017) indicamos que los sustantivos que lo ocupan prototípicamente «se hallan en el ámbito general de la 'comunicación verbal' o de la 'expresión corporal'», añadiendo que «el referente del sustantivo se particulariza a través de dos dimensiones diferentes y en ocasiones paralelas: la que alude a la forma de manifestarse el contenido del referente [...] y la que designa el contenido del referente» (2017: 255-256). El análisis más pormenorizado que ahora hemos llevado a cabo nos permite afinar un poco esta afirmación.

Por un lado, observamos que la forma de manifestación del referente se evidencia, en general, en todos los sustantivos, lo que no sucede en la dimen-

sión que podemos identificar como «semántica del referente»; en otras palabras, todos los sustantivos —en general— son susceptibles de clasificarse como distintas formas de manifestación del referente (véase *infra*), mientras que no todos expresan un sentimiento o estado de ánimo o bien una apreciación o un juicio desde uno de los dos polos, el positivo o el negativo, ya que hay sustantivos neutros, susceptibles de emplearse en ambos sentidos (*cfr.*: *gritos alegres*, *de denuncia*, *de dolor*, *de gozo*, *histéricos*…); la semántica del referente, por tanto, se establece en muchas ocasiones contextualmente.

Por otro lado, si bien en la mayor parte de los ejemplos encontrados de la construcción, las actualizaciones léxicas del *slot* sustantivo ofrecen una semántica relacionada con la emisión de sonidos (articulados y no articulados) o con gestos/mímica (en alusión a estados emocionales), hemos hallado un pequeño grupo de 54 registros, al que hemos aludido más arriba, de los campos semánticos 'beber', 'comer', 'fumar' —p. ej. *bocadillos*, *caladas*, *copas*[8]— que pueden, sin duda, interpretarse funcionalmente como instancias de la construcción a pesar de no cumplir la actualización del *slot* con las condiciones léxicas señaladas. Así, en el ejemplo (22), parafraseable con el significado de la construcción ('Fue allí, *bebiendo* tazas de té y gin-tonics, […]'), es evidente que la frase preposicional no indica localización física, sino que a través de una metonimia el objeto pasa a designar la acción que típicamente se hace con él.

Asimismo, hemos detectado un grupo de sustantivos situados en la dimensión del pensamiento: *admiraciones*, *delirios*, *dudas*, *incertidumbres*, *miedos*, *nervios*, *pensamientos*, *preocupaciones*, *sorpresas*, *titubeos*, *vacilaciones*… Salvo excepciones, también están marcados positiva o negativamente.

Creemos que en ambos casos nos situamos en la periferia de la construcción, lo cual viene avalado por el hecho de constituir un número muy reducido de sustantivos que, además, presentan frecuencias de uso muy bajas, así como por sus preferencias combinatorias, que no coinciden con las del resto de sustantivos: como ya hemos indicado y veremos con más detalle en el apartado 4.3, la clase semántica verbal con la que se combina más frecuentemente [*entre* + $S_{plural/corporal}$] es la de los verbos de comunicación y, secundariamente, también la de los verbos de desplazamiento; frente a esta preferencia general, los sustantivos de los campos semánticos 'beber', 'comer', 'fumar' y los que se sitúan en el ámbito del pensamiento presentan frecuencias de uso

[8] Prueba de que los sustantivos de estos ámbitos deben ser considerados como pertenecientes a la construcción es el hecho de que pueden ejercer coerción con el sustantivo que les sigue, como en *reflexionar entre botellas de vino y palos de golf*, con el significado 'reflexionar bebiendo vino y jugando al golf'. Si quitamos el primer elemento (*reflexionar entre palos de golf*) se bloquea el significado construccional y se activa el sentido local de la preposición.

sensiblemente más bajas con estas clases verbales y, por el contrario, son más proclives que el resto a combinarse con verbos de otras clases, como los de proceso existencial de fase-tiempo (*pasar el día entre bollos preñaos y sidra*) o los de proceso mental (*dirimir [algo] entre nervios*). A ello hay que sumar el hecho de que, en el caso del primer grupo, apenas se contabilizan registros que se adecuen formalmente al prototipo de la construcción (es decir, de sustantivo escueto sin modificación y sin coordinación, tipo *entre sorbos* o *entre cigarrillos*).

Teniendo en cuenta que nuestro objetivo principal es, precisamente, describir tal prototipo, hemos decidido eliminar estos registros de los recuentos. Dejando, entonces, a un lado estos dos grupos, así como el sustantivo *muestras*, cuyo núcleo denotacional se sitúa en su modificador (*entre muestras de agradecimiento*, *de cariño*, *de dolor*, *de nerviosismo*…), los 84 sustantivos con una frecuencia *token* mayor de 2, que constituyen aproximadamente el 30% de la muestra, se distribuyen en las distintas dimensiones del siguiente modo:

Tipo semántico de sustantivo	**Ejemplos**	**Nº de sust. con frecuencia *token* >2**	**Nº total de actualizaciones**
Discurso articulado	*alabanzas, ovaciones, acusaciones, insultos, juramentos, protestas*	28	188
Gritos y/o ruidos	*risas, abucheos, lloros, sollozos, gritos de dolor*	21	588[9]
Acciones con órganos o partes del cuerpo	*lágrimas, arcadas, convulsiones, escalofríos, espasmos*	18	217
Movimientos corporales, mímica, gesticulación	*abrazos, aplausos, arrumacos, caricias, sonrisas, aspavientos*	13	146
Total		**80**	**1139**

TABLA 1. Distribución del *slot* sustantivo por grupos semánticos y frecuencia *token*

1. Discurso articulado: 28 sustantivos con frecuencia *token* superior a 2 (el 35% de 80). Solo la mitad de ellos está marcada positiva o negativamente, de un modo, además, poco proporcional: apenas cuatro designan sentimientos o juicios positivos: *alabanzas*, *ovaciones*, *parabienes* y *vítores*, mientras que son diez los que se sitúan en el polo negativo: *acusaciones*, *amenazas*, *críticas*, *imprecaciones*, *insultos*, *juramentos*, *lamentos*, *maldiciones*, *protestas* y *quejas*. Los demás sustantivos son neutros respecto a la segunda dimensión,

[9] Hay que tener en cuenta que, de estas 588 ocurrencias, casi la mitad (274) corresponde a *entre risas*.

de modo que o no la expresan (*el tiempo pasa entre conversaciones*) o esta se extrae del contexto (*cfr.*: *entre comentarios admirativos / escandalizados*).

2. Gritos y/o ruidos, algunos generados como expresión de un estado emocional: 21 sustantivos (26,25%). Aunque algunos están marcados negativamente (*abucheos*, *lloros*, *sollozos*), en general el estado emocional se infiere del contexto (*gritos de dolor*, *de euforia*, *de gozo*, *histéricos*, *nerviosos*…) y, si bien suele predominar el sentido negativo, el sustantivo que presenta, con mucha diferencia, mayor frecuencia de uso (*risas*), casi siempre expresa un estado de ánimo positivo: «comunicarse entre risas»; «desplazarse entre risas»; «realizar, en general, alguna tarea entre risas» son las combinaciones más usuales de CORPES XXI. Apenas detectamos casos del tipo de *entre risas maliciosas*, *entre risas soeces*, *entre risas diabólicas*.

3. Acciones con órganos o partes del cuerpo: 18 sustantivos (22,5%). Como en el caso anterior, por lo general su significado no contiene la expresión de un sentimiento o un juicio positivo o negativo, pero los datos del corpus revelan que se emplean más con una carga negativa. El sustantivo más frecuente de este grupo, *lágrimas*, suele expresar un sentimiento de tristeza o desazón; de hecho, en el corpus no se registra ningún caso de *entre lágrimas de alegría*, pero sí *entre lágrimas de rabia y desesperación* o *entre lágrimas de dolor*. En general, los sustantivos que denotan movimientos involuntarios del cuerpo expresan malestar físico asociado a una situación adversa o un estado de ánimo desagradable: entre *arcadas*, *convulsiones*, *escalofríos*, *espasmos*, *hipos*, *sacudidas*, *sudores*, *temblores*, *vómitos*…

4. Movimientos corporales, mímica, gesticulación: 13 sustantivos (16,25%). Frente a los anteriores, en este grupo son más los sustantivos que están marcados positivamente: *abrazos*, *aplausos*, *arrumacos*, *caricias*, *sonrisas*, *bailes*; frente a *aspavientos* y *pucheros*.

Los datos cuantitativos y cualitativos obtenidos acerca de las actualizaciones del *slot* sustantivo de la construcción nos permiten llevar a cabo algunas consideraciones sobre la fijación cognitiva de la misma.

4.2. Productividad de la construcción de acuerdo con la frecuencia *type* y *token*

Los datos arriba indicados en relación con la diversidad de sustantivos que actualizan el *slot* en cada grupo semántico no indican en realidad preferencias en el empleo de un tipo semántico determinado. En efecto, este dato no ha de buscarse tanto en el número de sustantivos diferentes, como en el total de actualizaciones por grupo semántico. De hecho, solo en el último tipo de

sustantivos coinciden las tendencias entre unos datos y otros, es decir, el número de sustantivos con una frecuencia *token* > 2 es bajo (13 de 80), al igual que el número total de sustantivos (146 de 1139). Sin embargo, el grupo más nutrido de sustantivos con frecuencia *token* > 2, el de los que denotan discurso articulado, solo llega a las 188 ocurrencias, frente a las 217 de los que hacen referencia a actividades corporales y, sobre todo, las 588 actualizaciones de sustantivos que designan gritos y/o ruidos.

Ello es así, fundamentalmente, porque en la muestra se observa un gran desequilibrio en el empleo de los sustantivos: hemos registrado 297 sustantivos distintos, de los cuales 162 (algo más del 54% del total) se contabilizan solo una vez y 48 (en torno al 16%) se contabilizan dos veces. De hecho, únicamente hay 45 sustantivos que registren más de 5 apariciones (15%), 18 que registren más de 10 (6%) y 8 que registren más de 25 (cerca del 3%). Estas cifras indican que, aunque el *slot* pueda ser cubierto por una gran cantidad de sustantivos, el corpus revela una tendencia muy acusada a emplear frecuentemente un grupo pequeño de ellos. Incluso entre los más empleados, las cifras de frecuencia son muy dispares, oscilando entre los 20 registros de *gemidos* y *suspiros* y los 274 registros de *risas*. La siguiente tabla recoge las cifras de los sustantivos de uso más frecuente en el CORPES XXI. En la columna de la derecha se ofrece el número total de registros de cada sustantivo y en las centrales se desagrega esta cifra en los tres tipos sintácticos: sustantivo escueto (*entre lágrimas*), sustantivo expandido por subordinación (*entre lágrimas incontenibles*) y construcción coordinada (*entre gritos y lágrimas*):

	ESCUETO	EXPANDIDO	COORDINADO	Nº reg.
risas	206	19	49	274
lágrimas	83	11	9	103
sollozos	71	4	8	83
aplausos	21	12	21	54
gritos	6	27	19	52
bromas	28	3	12	43
carcajadas	22	4	2	28
sonrisas	14	2	11	27
susurros	13	3	6	22
gemidos	12	4	4	20
suspiros	16	1	3	20
	492	90	144	726

Tabla 2. Datos de frecuencia *token* de los once sustantivos más usados en CORPES XXI

Los once sustantivos de la tabla 2 abarcan aproximadamente la mitad de la muestra (726 registros), mientras que la otra mitad de la misma (709 regis-

tros)[10] se reparte entre los 286 sustantivos restantes. Lo que estos datos nos indican es que se puede perfilar fácilmente un núcleo prototípico de la construcción, constituido por sustantivos que manifiestan físicamente estados de ánimo polarizados, lo cual contribuye a que esta se fije cognitivamente[11].

Desde el punto de vista de la CxG, el hecho de que aparezca un elevado número de ocurrencias *hápax* (54% del total de los 297 sustantivos distintos) o con una frecuencia absoluta muy baja es un índice de la fijación de la construcción, esto es, a mayor diferenciación en la actualización de los *slots*, mayor fijación cognitiva (*entrenchment*) (*vid*. Ziem y Lasch 2013: 106). Por otro lado, este rasgo incide directamente en su grado de productividad. En palabras de Boas (2013: 247), la productividad «is measured in the context of a construction to determine how many different items occur in the various schematic slots of a construction».

Asimismo, también observamos tendencias a la fijación de determinados *tokens*, como *risas*, y, a larga distancia de este, *lágrimas* y *sollozos*, lo que podría apuntar a una cierta lexicalización del primero (*risas*), teniendo en cuenta que esta realización del *slot* sustantivo de la construcción —ya sea en forma escueta, modificada o coordinada— representa casi el 20% de los registros totales. No obstante, a diferencia de lo que sucede en otras construcciones en las que un *token* sobresale numéricamente sobre el resto (p. ej. *paso a paso* en la construcción [$S1_{sing}$ *a* $S1_{sing}$]), no se detecta ni idiomatización ni dispersión de significado en forma de polisemia o variantes semánticas (*cfr*. Iglesias Iglesias *et al*. 2018). En efecto, *entre risas* conserva un núcleo sémico acorde con el significado general de la construcción. Es, sin embargo, en su actualización coordinada, especialmente en la combinación más usual *entre risas y lágrimas*, donde se vislumbra quizás una cierta idiomatización hacia el significado 'en estado emocional confuso de alegría y tristeza'[12].

[10] La falta de concordancia en los datos ofrecidos —la suma de 726 y 709 (1435) no coincide con el número de registros indicado al principio de este apartado (1169)— se explica por el diferente modo de recuento en cada caso. Mientras que la cifra de 1169 corresponde a las instancias de la construcción contabilizadas en el CORPES XXI, la de 1435 alude a los sustantivos totales que hacen aparición en el conjunto de realizaciones, teniendo en cuenta que algunas de ellas contienen más de un sustantivo.

[11] La fijación cognitiva se observa ya en los testimonios del CORDE desde los siglos XVI y XVII, si bien la frecuencia *token* no coincide del todo con la que revela el corpus sincrónico CORPES XXI: como instancia más usual aparece *entre sollozos* (129 registros), seguida de *entre lágrimas* (127), *entre risas* (109), *entre gritos* (43) y *entre aplausos* (33). Llama la atención la coincidencia en el tipo semántico de las realizaciones léxicas del *slot* con el presente y sorprende la alta incidencia de la modificación por adjetivo y la coordinación, a excepción, curiosamente, de *sollozos*, que revela la frecuencia *token* más alta y por ello se le supone un mayor grado de fijación cognitiva.

[12] Se han realizado búsquedas para analizar si el binomio coordinado *entre risas y lágrimas* ofrece algún tipo de fijación en el orden de sus constituyentes, pero no se ha podido verificar nuestra intui-

Respecto a los sustantivos que se contabilizan solo una vez (162), solo 23 corresponden a la construcción con sustantivo escueto y, en general, coinciden exactamente con los grupos que hemos señalado: *entre resuellos*, *risillas*, *ronroneos*, *ronquidos*, *tarascadas*, *tumbos*, etc. Otros 28, también muy pocos, encajan con la construcción expandida por subordinación y también, en general, corresponden a los tipos esperados: *entre balidos lastimeros*, *entre burlas cariñosas*, *entre grandes quejidos*, *entre ebrios forcejeos*, *entre extrañas contorsiones*, etc. La mayoría, sin embargo, aparece de forma coordinada (111) y, aunque se detectan muchos sustantivos perfectamente adecuados a la caracterización ofrecida (*agasajos*, *aleluyas*, *aprobaciones*, *carrasperas*, *disculpas*, *lametones*, etc.), también hay muchos otros que no lo son: *amoríos*, *banderas*, *desquites*, *errores*, *proyectos*, etc.[13]

Al igual que sucede con sustantivos encontrados en las realizaciones de la construcción, del tipo *gestos*, *muestras* o *palabras* (*entre gestos de escepticismo*, *entre muestras de júbilo*, *entre palabras de despedida*) (*vid. supra*), entre los *tokens* más frecuentes también hallamos algunos que tienden por su significado a un uso con modificador o en coordinación, y no tanto como sustantivo escueto, como es el caso de *gritos* (6 escueto frente a 46 modificado/coordinado). También *aplausos* iguala en su uso coordinado (21) al escueto (21), lo que llama la atención, ya que tanto las actualizaciones de la construcción con *gritos* como con *aplausos* se corresponden al tipo 2. Con todos estos datos podríamos formular un prototipo de construcción con sustantivo escueto de tipo 1, que coincide con los de una frecuencia *token* elevada, de tal manera que a menor frecuencia *token*, mayor distanciamiento del prototipo y mayor tendencia a la expansión con modificadores y a la coordinación.

Los datos aportados nos permiten hacer algunas consideraciones acerca de la fijación cognitiva de la construcción (*entrenchment*) y la productividad de esta, ambas en estrecha relación. La fijación cognitiva puede referirse al *type* de la construcción, vista esta como una fórmula abstracta de emparejamiento entre una forma y un significado. Y la fijación cognitiva también puede aludir a los distintos *tokens* o actualizaciones por separado.

---

ción, dado que el número de registros de *entre lágrimas y risas* presenta igualmente una alta frecuencia. Para esta búsqueda se ha acudido al corpus Sketch Engine *Spanish Web 2011 es TenTen11 EU+AM*, pues CORPES XXI apenas ofrecía registros.

[13] La presencia de estos sustantivos en la muestra tiene, en nuestra opinión, dos explicaciones, ya señaladas: pueden encajar en la construcción a través del principio de coerción (véase ej. (14) *entre gritos, consignas y pancartas*) o bien deberse a relaciones de herencia de construcciones más esquemáticas con *entre* (véanse los ejs. (17) *entre brincos y desenfreno* y (18) *entre cirios y susurros*). También a relaciones interconstruccionales, en este caso con [*entre* S1 *y* S1], si bien estos últimos sustantivos no han sido incluidos en los recuentos, como ya señalamos (véase ej. (22) *entre tazas de té y gin-tonics*).

La fijación cognitiva del *type* está directamente vinculada con la frecuencia absoluta de instancias que actualizan la construcción (Goldberg 2006: 5) y por extensión con el número de las distintas realizaciones de los *slots*: a mayor variedad de actualizaciones de los *slots*, mayor productividad (Bybee 2013: 61) y por ende mayor fijación. Además de este criterio, Ziem y Lasch (2013: 106) citan el argumento del número de realizaciones únicas de cada *slot*, es decir, de *hápax*.

En cuanto a la fijación cognitiva de un *token*, esta es proporcional a su frecuencia absoluta[14]. El estudio de la frecuencia *token* permite ver el grado de apertura de la construcción, es decir, la variabilidad de los ítems de un determinado patrón (Boas 2013), así como la llamada *statistical pre-emption* (Goldberg 2006: 93), referida a la presencia cuantitativa de un determinado *token* en relación con una construcción cercana[15].

Con respecto a nuestra construcción, destacamos a modo de resumen los siguientes datos, mencionados más arriba:

Datos de frecuencia para el *type-entrenchment*:

i. 1169 registros del CORPES XXI que se ajustan al prototipo de la construcción [*entre* + $S_{plural/corporal}$].
ii. Número total de actualizaciones del *slot* sustantivo (contando la coordinación): 1435.
iii. Número de sustantivos distintos: 297.
iv. Número de *hápax legómena* entre las actualizaciones del slot sustantivo: 162 (54%).
v. Datos de frecuencia para el *token-entrenchment* (11 actualizaciones más frecuentes): *risas* (274), *lágrimas* (103), *sollozos* (83), *aplausos* (54), *gritos* (52), *bromas* (43), *carcajadas* (28), *sonrisas* (27), *susurros* (22), *gemidos* (20), *suspiros* (20).

[14] Un aspecto muy interesante es determinar a partir de cuándo un *token* de una construcción se convierte en un *type*, es decir, en una construcción propia, lo cual depende sobremanera de su fijación cognitiva y su capacidad para generar nuevas instancias (*cfr*. Mollica & Schafroth 2018: 131). Esto sería lo que sucede cuando una unidad fraseológica sufre una fuerte variación de un constituyente en el discurso (*cfr*. Mellado Blanco 2018b a propósito del refrán alemán *Reden ist Silber, Schweigen ist Gold*).

[15] Estos dos últimos datos no serán analizados en el presente trabajo por sobrepasar los objetivos iniciales del mismo. En este punto, sería interesante analizar la frecuencia *token* de la construcción con coordinación [*entre* + $S_{plural/corporal}$] y la construcción [*entre* $S1_{sing}$ *y* $S1_{sing}$]. De acuerdo con los estudios realizados en López Meirama (2017) y en el presente trabajo, podemos afirmar de manera genérica que el prototipo de *tokens* de la construcción [*entre* + $S_{plural/corporal}$] no es compartido por la construcción [*entre* $S1_{sing}$ *y* $S1_{sing}$] y viceversa. De hecho, las actualizaciones prototípicas *risa, lágrima, sollozo* de [*entre* + $S_{plural/corporal}$] son *hápax* en la segunda de acuerdo con las calas realizadas en CORPES XXI.

Resumiendo, los datos aportados se traducen en un alto grado de fijación cognitiva de la construcción (*type-entrenchment*) debido a la gran variedad de actualizaciones distintas del *slot* sustantivo y de un índice de más del 50% de *hápax*. A su vez se distingue un núcleo prototípico de actualizaciones *token* (11 sustantivos en total), que abarca aproximadamente la mitad de los registros de la construcción extraídos del corpus. Dichos sustantivos presentan gran proximidad semántica entre sí, lo que apunta a un significado nuclear de la construcción, que denota expresivamente estados de ánimo polarizados simultáneos a la acción descrita por el verbo principal (colocado), que, como veremos en 4.3, es con frecuencia de comunicación. Entre los sustantivos con mayor representación cuantitativa en el corpus, hemos visto ya que destaca *risas* de manera clara.

4.3. Datos sobre los colocados de la construcción: las clases verbales

Hemos realizado un análisis detallado del corpus para averiguar si se detecta o no alguna tendencia en las clases semánticas de los verbos con los que suele coaparecer [*entre* + $S_{plural/corporal}$]. El resultado es que, si bien la construcción puede combinarse con cualquier tipo de verbo, se manifiesta una muy clara preferencia, en primer lugar por los verbos de comunicación: *admitir*, *afirmar*, *asegurar*, *comentar*, *confesar*, *contar*, *decir*, *explicar*, *hablar*, *preguntar*, *reconocer*, *responder*, etc. y, en segundo lugar, por los verbos de desplazamiento: *abandonar (un lugar)*, *acercarse*, *caer*, *correr*, *entrar*, *ir(se)*, *llegar*, *llevar*, *regresar*, *salir*, *subir*, *volver*, etc.[16]

Para llevar a cabo el análisis, hemos partido de la clasificación verbal de la *Base de datos de Verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español* de la Universidad de Vigo (en adelante, ADESSE), si bien la hemos adaptado para nuestros propósitos en los siguientes extremos:

1. Hemos clasificado como verbos de comunicación todos los que en nuestro corpus funcionan como tales, aunque no se hayan catalogado así en ADESSE, como *consolar*, caracterizado en esta base de datos como verbo MENTAL: *La consoló entre susurros al oído*. En muchos casos, la clasificación se justifica porque el verbo introduce una secuencia de estilo directo, como en el caso de *resaltar*, que ADESSE clasifica en el grupo EXISTENCIAL:

(24) «Si no fuera por lo que es, no estaban ahí para meternos prisa», resalta, *entre sonrisas,* ante la distraída mirada de Rafael Losa. (M. Balín: «Obras a pie de La Almudena». *Diario de León.es*.)

[16] Este resultado corrobora la descripción ofrecida en Mellado Blanco & López Meirama (2017).

2. Hemos considerado la existencia de «predicados complejos»: *desgranar desgracias* y *hacer una declaración*, por ejemplo, han sido incluidos entre los verbos de comunicación; *hacer el viaje de regreso*, entre los de desplazamiento.

3. Algunos de los verbos de nuestro corpus no están registrados en ADESSE: *expirar* (clasificado como de proceso EXISTENCIAL - clase 'vida'), *desbastar* (MATERIAL-'cambio'), *pifiar* (MENTAL-'cognición'), *requebrar* (VERBAL-'valoración').

Los seis tipos de proceso recogidos en ADESSE son los siguientes (adjuntamos la caracterización que de cada uno se ofrece en la Base de datos, así como algunos ejemplos):

1. Tipo de proceso: MENTAL: 88 casos.

«Una entidad dotada de vida psíquica (A1) mantiene o experimenta algún tipo de estado, cambio de estado o actividad interior perceptiva, sensitiva y/o cognitiva (A2)»: *comprender entre lágrimas de emoción*, *imaginar entre escalofríos*, *escuchar entre lloros*.

2. Tipo de proceso: RELACIONAL: 28 casos.

«Macroclase que incluye sobre todo verbos de atribución y de posesión»: *servir para algo entre risas de incredulidad y cariño*, *ser algo entre engaños y consuelos*, *denegar algo entre risas*.

3. Tipo de proceso EXISTENCIAL: 86 casos.

«Clase genérica que incluye verbos de existencia, tiempo-fase y vida»: *acabar algo entre vítores y aplausos*, *morir entre espasmos*, *surgir algo entre ahogos*. En el caso de sustantivos relacionados con el *dolor*, destaca la combinabilidad de estos con verbos existenciales (*morir*, *acabar sus días*, *nacer entre entre fuertes/terribles/horribles dolores*).

4. Tipo de proceso: MODULACIÓN: 15 casos.

«Macroclase que incluye las clases con verbos próximos a los gramaticalizados como auxiliares o semiauxilares: causativos, dispositivos, verbos soporte»: *ceder a algo entre sollozos*, *fomentar algo entre burlas cariñosas*, *negarse a algo entre lloros*.

5. Tipo de proceso: MATERIAL: 404 casos.

«Macroclase que incluye distintas subclases de procesos físicos (no mentales)». Este grupo se divide en las clases 'cambio' (*mojarse entre carcajadas*), 'comportamiento' (*comer entre sonrisas*) y 'espacio' (*esconderse entre risas*). Dentro de la clase 'espacio', que suma 224 casos, hay 173 que corres-

ponden a la subclase 'desplazamiento': *acercarse entre bostezos lastimeros*, *correr entre grandes risotadas*, *desembarcar entre aplausos y vítores*.

6. Tipo de proceso: VERBAL: 548 casos

«Clase genérica que incluye verbos de comunicación, valoración y emisión de sonido» [17]. Estos últimos (*jadear entre escalofríos de horror*) pueden distinguirse del resto, que constituyen la tradicional clase de los *verba dicendi* (verbos de comunicación: *hablar entre sollozos*; de petición: *suplicar entre llantos y gritos*; de valoración: *abuchear entre risas*).

Los datos cuantitativos evidencian que de los seis tipos de proceso destacan claramente dos, el VERBAL y el MATERIAL. Además, entre los verbos de proceso MATERIAL predomina el grupo de los verbos de 'espacio'; más concretamente, el subgrupo de verbos de desplazamiento. El gráfico 2 permite visualizar estas preferencias (en los procesos MATERIAL y VERBAL hemos marcado en claro las clases más frecuentes: en el tipo de proceso MATERIAL, la subclase 'desplazamiento'; en el verbal, las clases 'comunicación' —incluye 'petición'— y 'valoración', dejando fuera la clase 'emisión de sonido', apenas representada en nuestro corpus con cinco registros).

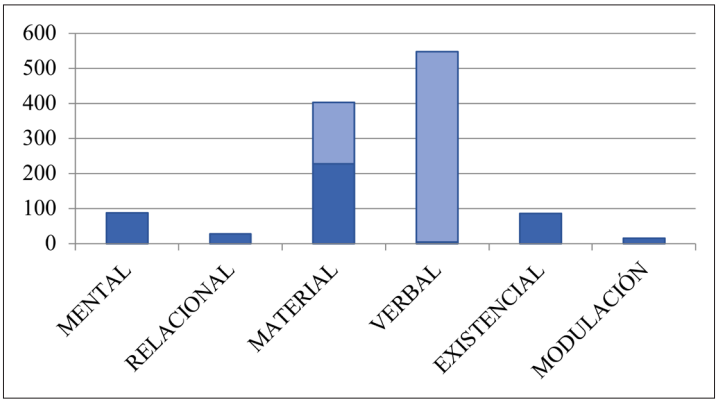


GRÁFICO 2. Preferencias en la combinatoria verbal

Estos datos se recogen en la tabla y el gráfico siguientes:

[17] El estudio de los verbos de comunicación y su tipología construccional ha sido enfocado desde la CxG y de la Lingüística Cognitiva en varias publicaciones, tanto sobre el español (*cfr*. Martínez Vázquez 2005) como sobre el inglés (Faber & Sánchez 1990 y Boas 2010). Por motivos de espacio no podemos ahondar aquí en este tema.

Verba dicendi	548
Verbos de desplazamiento	173
Otros verbos	448

TABLA 3. Distribución de las clases más frecuentes

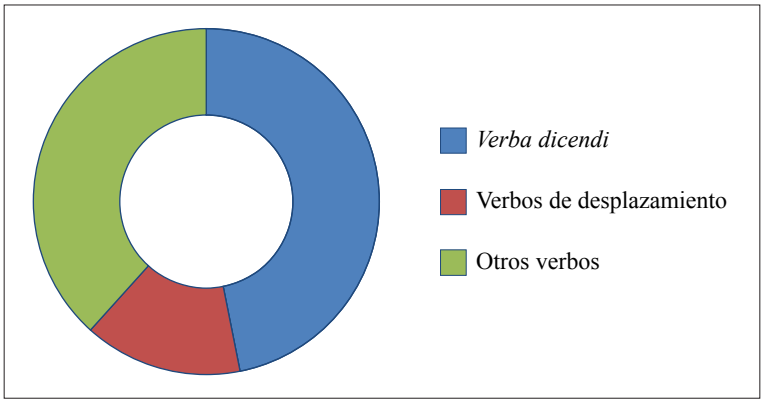


GRÁFICO 3. Distribución de las clases más frecuentes

Los verbos de comunicación constituyen, pues, el grupo principal de verbos que coocurren con la construcción estudiada. Llama la atención, igualmente, el elevado uso del estilo directo, como corrobora la siguiente tabla, en la que se recogen las subclases de este grupo de verbos:

Subclase verbal	**Nº tokens**	**Nº tokens ED**	**% ED**
Comunicación	510	264	51,76%
Petición	23	1	4,35%
Valoración	10	0	0%
Emisión de sonido	5	0	0%

Tabla 4. Número de *tokens* de la construcción según la subclase verbal y porcentajes de estilo directo

En estos casos, la secuencia de discurso directo depende directamente del verbo de comunicación, como se observa en el ejemplo (25) con *sonsacar*:

(25) — Dicen de ti que vas a comerte el mundo, ¿es eso cierto? —le sonsaco *entre besos y bocados*, ronroneando como una gata zalamera complacida por las caricias de su amo. (M. Castro: *Mantis*)

En otros casos, sin embargo, el discurso directo depende del sustantivo del tipo *grito*, *voz*, *coro*, *denuncia*, *crítica*, etc., como en el ejemplo (26):

(26) «[…] hace falta es que la gente de [*sic*] un gran castigo a Aznar en las próximas elecciones», ha señalado Saura en un punto en el que ha coincidido con todos los ponentes que han hablado a los manifestantes *entre gritos de «No a la guerra» y «No a la guerra, no al PP»*. («Miles de personas protestan en Barcelona contra el conflicto en Irak». *El Mundo.es*)

Como sucedía con la frecuencia *token* del slot *sustantivo*, también en el caso de los verbos que se combinan con la construcción las preferencias en el uso están muy marcadas: de los datos extraídos del corpus se desprende que, mediante la elección de [*entre* + $S_{plural/corporal}$], el hablante quiere resaltar de manera expresiva la circunstancia emotiva que acompaña al agente del enunciado, básicamente mientras emite un mensaje (caso de combinatoria con *verba dicendi*) (27) o, en menor medida, al realizar una acción que implica desplazamiento (caso de combinatoria con verbos de desplazamiento) (28):

(27) En un tono de complicidad amistosa *le confiesa entre risas* que se alegra de que Tomás Campuzano se haya llevado todos los muebles del despacho. (M. Tena: *Tenemos que vernos*)

(28) Dora *avanzaba* lenta hacia su casa *entre sollozos* sin que a nadie le importase. (F. Casavella: *Los juegos feroces*)

En los ejemplos se constata que tal «circunstancia emotiva» suele ser una actividad llevada a cabo por el propio agente del enunciado, por lo que, como ya hemos comentado, la construcción funciona de un modo similar a como lo hace un gerundio (*Se lo confiesa riéndose; Avanzaba sollozando*), si bien no siempre es así: en el corpus también hemos detectado ejemplos como los siguientes, en los que el agente (o, a veces, el experimentador) de la actividad denotada metonímicamente por el sustantivo no coincide con el del enunciado (véanse también los ejemplos (11) y (12), en los que se explicita el agente mediante una frase preposicional adyacente del sustantivo):

(29) Fernando Trueba intentó explicar, *entre abucheos,* que se puede estar en contra de la guerra y en contra de Fidel. (P. Ordaz: «Un acto contra Castro acaba en un ataque furibundo al PSOE». *El País*)

(30) Tras el tanto visitante, Aguirre sacó del terreno de juego a Eller, que lo abandonó *entre silbidos,* y dio entrada a Luis García («El Atlético no supera su depresión». *Superdeporte.es*)

Los casos ilustrados en (29) y (30) son minoritarios, ya que apenas alcanzan el 20% de los 1169 registros que componen la muestra extraída del CORPES XXI, lo cual, en nuestra opinión, demuestra cuantitativamente el carácter menos nuclear de estas realizaciones. La distribución de los datos por clases semánticas, además, avala este punto de vista. En la tabla 5 se ofrecen las cifras desagregadas por las clases semánticas verbales.

Tipo de proceso	**Subtipo 1** **Mismo actante**	**Subtipo 2** **Distinto actante**
verbal	515	33
desplazamiento	111	62
material no desplazamiento	181	50
mental	69	19
existencial	42	44
relacional	13	15
modulación	12	3
TOTAL	943	226

TABLA 5. Distribución mismo actante / distinto actante por clases semánticas verbales

Las cifras de la tabla 5 evidencian un comportamiento no homogéneo de las clases verbales, que se puede observar más fácilmente en el siguiente gráfico (para visualizar mejor las diferencias, se marca con un trazo vertical en azul el porcentaje general de los casos con el mismo actante, aproximadamente el 80% del total). Debe tenerse en cuenta, no obstante, que el gráfico ofrece datos porcentuales, no frecuencias absolutas, y que las diferencias en estas últimas son muy grandes entre unas clases verbales y otras.

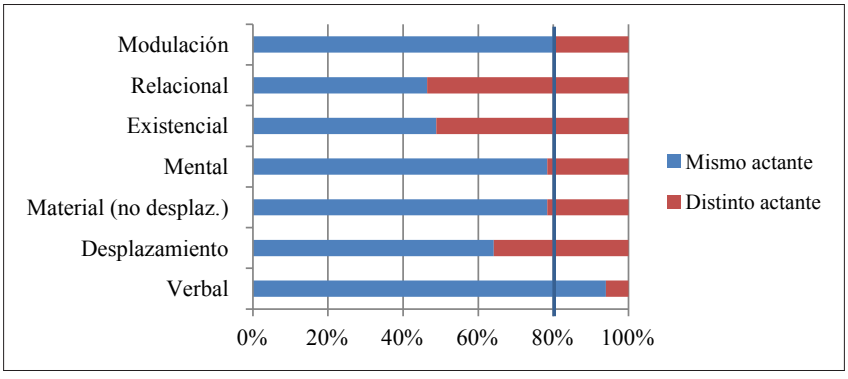


GRÁFICO 4. Distribución porcentual mismo actante / distinto actante por clases semánticas verbales

En buena medida, las diferencias se explican por la propia semántica verbal, al menos en lo que se refiere a los verbos de proceso relacional y de proceso existencial: en el primer grupo, el actante suele coincidir cuando el verbo pertenece a la clase 'Posesión' (*devolver algo entre risas*, *regalar algo entre lágrimas*) y suele no hacerlo si es de 'Atribución' (*ser algo entre risas*); en el segundo, los verbos de la clase 'Fase-tiempo' son los que marcan la diferencia: los acontecimientos se desarrollan, continúan, transcurren, pasan, etc., mientras alguien hace algo: *El día transcurre entre bromas*, *El encuentro termina entre aplausos*, etc. Dentro del grupo de verbos de tipo 'Existencial' (existencia, fase-tiempo y vida), llama precisamente la atención el bloque de verbos fase-tiempo (representa el 62% del grupo 'Existencial'), lo cual se encuentra en relación directa con el significado de marco temporal de nuestra construcción, como se ve en (20).

Por otra parte, el porcentaje tan elevado de coincidencia de actantes en los verbos más destacados del corpus, los de proceso verbal, apunta a la existencia de un prototipo de la construcción que está fijado cognitivamente: alguien se comunica [de determinada manera]. Esa manera nos indica la actitud del hablante en el momento de la comunicación: o bien informa de su estado de ánimo (*Me lo dijo entre lágrimas*) o de su punto de vista (*Daba instrucciones a gritos y entre imprecaciones*).

5. A MODO DE BALANCE

El estudio basado en corpus se confirma como un método fiable y consistente para estudiar el funcionamiento y la fijación cognitiva de las construcciones. En este trabajo hemos optado por una construcción preposicional, parcialmente esquemática, [*entre* + $S_{plural/corporal}$], que hasta ahora ha pasado prácticamente desapercibida en las gramáticas en su calidad de predicación secundaria y en sus rasgos semántico-pragmáticos. El análisis en CORPES XXI España arroja interesantes resultados acerca de sus realizaciones prototípicas, tanto en lo que se refiere a las actualizaciones del *slot* sustantivo (frecuencia *type* y *token*), como en relación con su combinatoria verbal, lo cual permite sugerir que esta construcción exhibe un alto grado de fijación cognitiva. Este hecho explicaría el alto índice de casos de coerción detectados. Asimismo, hemos constatado relaciones de herencia e interconstruccionales de la construcción respecto a otras del *contínuum* léxico-gramatical. En este sentido, [*entre* + $S_{plural/corporal}$] muestra en algunos ejemplos relaciones de herencia vertical, que se plasman en matices propios del significado local, causal o sumativo de *entre*. En otras ocasiones son evidentes las relaciones interconstruccionales con la también construcción modal-temporal de predicación

secundaria [*entre* $S1_{singular}$ *y* $S1_{singular}$], en especial cuando el *slot* sustantivo pertenece al campo 'comer', 'beber' y 'fumar' (*comentar entre copas / entre copa y copa*). Estas relaciones, en sentido vertical y horizontal, refuerzan la idea de interconexión entre las construcciones del constructicón de la lengua.

Corpus

ADESSE: Base de datos de verbos, alternancias de diátesis y esquemas sintáctico-semánticos del español. Universidade de Vigo. http://adesse.uvigo.es

Davies, Mark (2002-) *Corpus del español: 100 million words, 1200s-1900s*. http://www.corpusdelespanol.org

CORPES: Real Academia Española: Banco de datos (CORPES XXI) [en línea]. *Corpus del Español del Siglo XXI (CORPES)*. http://www.rae.es

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

Sketch Engine *Spanish Web 2011 es TenTen11 EU+AM*: Sketch Engine: Corpus Query System. http://www.sketchengine.eu/

Referencias bibliográficas

Boas, Hans C. (2010): «Linguistically relevant meaning elements of English communication verbs», *Belgian Journal of Linguistics* 24, pp. 54-82. https://doi.org/10.1075/bjl.24.03boa

Boas, Hans C. (2011): «Coercion and leaking argument structures in Construction Grammar», *Linguistics* 49/6, pp. 1271-1303. https://doi.org/10.1515/ling.2011.036

Boas, Hans C. (2013): «Cognitive Construction Grammar», in Thomas Hoffmann & Graeme Trousdale (eds.): *The Oxford handbook of Construction Grammar*. Oxford: Oxford University Press, pp. 233-252. https://doi.org/10.1093/oxfordhb/9780195396683.013.0013

Butler, Christopher S. & Francisco Gonzálvez-García (2014): *Exploring functional-cognitive space*. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.157

Bybee, Joan L. (2013): «Usage-based theory and exemplar representations of constructions», in Thomas Hoffmann & Graeme Trousdale (eds.): *The Oxford handbook of Construction Grammar*. Oxford: Oxford University Press, pp. 49-69. https://doi.org/10.1093/oxfordhb/9780195396683.013.0004

Cabezas Holgado (2013): *La predicación: las construcciones en abanico. Aplicaciones al español*. Tesis doctoral. Universidad Complutense de Madrid. http://eprints.ucm.es/22365/1/T34644.pdf

Croft, William (2001): *Radical Construction Grammar: syntactic theory in typological perpespective*. New York: Oxford University Press.

Faber, Pamela B. & Jesús Sánchez (1990): «Semántica de prototipos: el campo semántico de los verbos que expresan la manera de hablar frente al de los verbos de sonido en inglés y español», *Revista de Lingüística Aplicada* 6, pp. 19-29.

Finkbeiner, Rita (2015): «*Wie deutsch ist DAS denn?!* Satztyp oder Konstruktion?», in Charlotta Brylla & Elisabeth Wåghäll-Nivre (eds.): *Sendbote zwischen den Kulturen: Gustav Korlén und die germanistische Tradition an der Universität Stockholm*. Estocolmo: Acta Universitatis Stockholmiensis, pp. 243-273.

García-Page Sánchez, Mario (2007): «Esquemas sintácticos de formación de locuciones adverbiales», *Moenia: Revista lucense de lingüística e literatura* 13, pp. 121-144.

Goldberg, Adele (1995): *A Construction Grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, Adele (2006): *Constructions at work: the nature of generalizations in language*. Oxford: Oxford University Press.

González-García, Francisco (2013): «Breves consideraciones en torno a la interacción entre coerción y polisemia: el caso de la predicación secundaria con verbos de cognición en español», in Sabine De Knop, Fabio Mollica & Julia Kuhn (eds.): *Konstruktionsgrammatik in den romanischen Sprachen*. Berlín: Peter Lang, pp. 184-203.

González-García, Francisco (2014): «Words as constructions: some reflections in the light of constructionist approaches», in Iraide Ibarretxe-Antuñano & José Luis Mendívil Giró (eds.): *To be or not to be a word: new reflections on the definition of word*. Newcastle: Cambridge Scholars, pp. 164-188.

González-García, Francisco (2017): «Exploring inter-constructional relations in the constructicon: a view from Contrastive (Cognitive) Construction Grammar», in Francisco José Ruiz de Mendoza Ibanez, Alba Luzondo Oyón & Paula Pérez Sobrino (eds.): *Constructing families of constructions (human cognitive processing)*. Amsterdam: John Benjamins, pp. 135-172. https://doi.org/10.1075/hcp. 58.06gar

Hernández Díaz, Axel (2014): «Las preposiciones *en* y *entre*», in Concepción Company Company (dir.): *Sintaxis histórica de la lengua española. Tercera parte: Adverbios, preposiciones y conjunciones. Relaciones interoracionales*. Ciudad de México: Fondo de Cultura Económica – Universidad Nacional Autónoma de México, pp. 1631-1721.

Iglesias Iglesias, Nely M., Carmen Mellado Blanco & Ana Mansilla Pérez (2018): «El esquema fraseológico del alemán [X *für* X] y su(s) correspondencia(s) en español: acercamiento constructivista desde la lingüística de corpus». Ponen-

cia presentada en el Congreso Internacional de Traducción e Interpretación: «Traducción y sostenibilidad cultural», Facultad de Traducción y Documentación de la Universidad de Salamanca, 28-30 de noviembre de 2018.

Kiss, Tibor, Antje Müller, Claudia Roch, Tobias Stadtfeld, Katharina Börner & Monika Duzy Ein (2014): *Handbuch für die Bestimmung und Annotation von Präpositionsbedeutungen im Deutschen*. Bochum: Sprachwissenschaftliches Institut, Ruhr-Universität Bochum.

Lauwers, Peter (2010): «Les locutions en 'sous' comme constructions», *Le Français Moderne* 78/1, pp. 3-27.

Lauwers, Peter & Dominique Willems (2011): «Coercion: definition and challenges, current approaches, and new trends», *Linguistics* 49/6, pp. 1219-1235. https://doi.org/10.1515/ling.2011.034

López Meirama, Belén (2017): «*Entre trago y trago*: la construcción binomial [*entre* + S1 y S1] en español», in Silvia Gumiel-Molina, Manuel Leonetti & Isabel Pérez-Jiménez (eds.): *Investigaciones actuales en lingüística*. Vol. III: *Sintaxis*. Alcalá de Henares: Universidad de Alcalá, pp. 95-110.

Maldonado, Ricardo (2010): «*Claro*: de objeto perceptible a refuerzo pragmatico», in María José Rodríguez Espiñeira (ed.): *Adjetivos en discurso: emociones, certezas, posibilidades y evidencias*. Santiago de Compostela: Universidade. Servizo de Publicacións e Intercambio Científico, pp. 61-107.

Martínez García, Hortensia (2003): *Construcciones temporales*. Madrid: Arco Libros, 2ª edición.

Martínez Vázquez, Montserrat (2005): «Communicative constructions in English and Spanish», in Christopher Butler, María de los Ángeles Gómez González & Susana M. Doval-Suárez (eds.): *The dynamics of language use: functional and contrastive perspectives*. Amsterdam: John Benjamins, pp. 79-109.

Mellado Blanco, Carmen (2017): «*A la luz de los corpus*: semántica y análisis de coocurrencias de <*a la luz de* + SN>», in Claudia Zavaglia & Angélica Karim Garcia Simão (orgs.): *Reflexões, tendências e novos rumos dos estudos fraseoparemiológicos*. São José do Rio Preto: Universidade Estadual Paulista / Instituto de Biociências, Letras e Ciências Exatas, pp. 28-45.

Mellado Blanco, Carmen (2018a): «*Unter Tränen, unter Beifall*: Das Präpositionsmuster [*unter* + SUB_{SOMAT} (+ *von* / Genitivattribut)]», in Natalia Filatkina & Sören Stumpf (eds.): *Conventionalization and variation / Konventionalisierung und Variation*. Frankfurt am Main: Peter Lang, pp. 201-228.

Mellado Blanco, Carmen (2018b): «Wenn modifizierte Sprichwörter zu Mustern werden: Eine korpusbasierte Studie am Beispiel von *Reden ist Silber, Schweigen ist Gold*», in Martina Nicklaus, Nora Wirtz, Marcella Costa, Karin

Ewert-Kling & Wiebke Vogt (eds.): *Lexeme, Phrasen… Konstruktionen: Aktuelle Beiträge zur Lexikologie und Phraseologie*. Frankfurt am Main: Peter Lang, pp. 183-203.

Mellado Blanco, Carmen (en prensa): «Esquemas fraseológicos y construcciones fraseológicas en el *contínuum* léxico-gramática», in Encarnación Tabares, Carsten Sinner & Esteban T. Montoro (eds.): *Clases y categorías en la fraseología de la lengua española*. Frankfurt am Main: Peter Lang.

Mellado Blanco, Carmen & Belén López Meirama (2017): «Esquemas sintácticos de [PREP + S]: el caso de [*entre* + $S_{plural/corporal}$]», in Carmen Mellado Blanco, Katrin Berty & Inés Olza (eds.): *Discurso repetido y fraseología textual (español y español-alemán)*. Madrid / Frankfurt am Main: Iberoamericana / Vervuert, pp. 249-267. https://doi.org/10.31819/9783954876037-014

Michaelis, Laura A. (2003a): «Headless constructions and coercion by construction», in Elaine J. Francis & Laura A. Michaelis (eds.): *Mismatch: form-function incongruity and the architecture of grammar*. Stanford: CSLI Publications, pp. 259-310.

Michaelis, Laura A. (2003b): «Word meaning, sentence meaning, and syntactic meaning», in Hubert Cuyckens, René Dirven, John R. Taylor & Ronald W. Langacker (eds.): *Cognitive approaches to lexical semantics*. Berlín: Mouton de Gruyter, pp. 163-209. https://doi.org/10.1515/9783110219074.163

Michaelis, Laura A. (2004): «Type shifting in Construction Grammar: an integrated approach to aspectual coercion», *Cognitive Linguistics* 15, pp. 1-67. https://doi.org/10.1515/cogl.2004.001

Michaelis, Laura A. (2011): «Stative by construction», *Linguistics* 49/6, pp. 1359-1400. https://doi.org/10.1515/ling.2011.038

Mollica, Fabio & Elmar Schafroth (2018): «Der Ausdruck der Intensivierung in komparativen Phrasem-Konstruktionen im Deutschen und im Italienischen: eine konstruktionsgrammatische Untersuchung», in Kathrin Steyer (ed.): *Sprachliche Verfestigung: Wortverbindungen, Muster, Phrasem-Konstruktionen*. Tübingen: Narr Francke Attempto Verlag, pp. 103-136.

Montoro del Arco, Esteban T. (2008): «El concepto de locución con casillas vacías», in Carmen Mellado Blanco (ed.): *Colocaciones y fraseología en los diccionarios*. Frankfurt am Main: Peter Lang, pp. 131-146.

Pustejovsky, James (1991): «The syntax of event structure», *Cognition* 41, pp. 47-81. https://doi.org/10.1016/0010-0277(91)90032-Y

Ruiz Gurillo, Leonor (1998): «Una clasificación no discreta de las unidades fraseológicas del español», in Gerd Wotjak (ed.): *Estudios de fraseología y fraseografía del español moderno*. Madrid / Frankfurt am Main: Iberoamericana / Vervuert, pp. 13-37. https://doi.org/10.31819/9783865278371-002

Schafroth, Elmar (2013): «Das pragmatische Potential von Phrasemen – illustriert am Deutschen und Italienischen», in Sibilla Cantarini (ed.): *Wortschatz, Wortschätze im Vergleich und Wörterbücher: Methoden, Instrumente und neue Perspektiven*. Frankfurt am Main: Peter Lang, pp. 185-208.

Seco, Manuel, Olimpia Andrés & Gabino Ramos (2011): *Diccionario del español actual*. Madrid: Aguilar, 2a edición. [DEA]

Taylor, John R. (2016): «Cognitive Linguistics», in Keith Allan (ed.): *The Routledge Handbook of Linguistics*. Londres / Nueva York: Routledge, pp. 455-469.

Traugott, Elizabeth (2018): «Modeling language change with constructional networks», in Salvador Pons Bordería & Óscar Loureda Lamas (eds.): *Beyond grammaticalization and discourse markers*. Leiden: Brill, pp. 17-50.

Van de Velde, Freek (2014): «Degeneracy: the maintenance of constructional networks», in Ronny Boogaart, Timothy Colleman & Gijsbert Rutten (eds.): *Extending the scope of Construction Grammar*. Berlín: de Gruyter Mouton, pp. 141-180.

Ziem, Alexander & Alexander Lasch (2013): *Konstruktionsgrammatik: Konzepte und Grundlagen gebrauchsbasierter Ansätze*. Berlín: de Gruyter. https://doi.org/10.1515/9783110295641

Zuluaga, Alberto (1980): *Introducción al estudio de las expresiones fijas*. Frankfurt am Main: Peter Lang.

EN TORNO AL CONCEPTO DE *PERFIL COMBINATORIO*

On the concept of behaviroral profile

Inmaculada Mas Álvarez
Universidade de Santiago de Compostela

Laß dich die Bedeutung der Worte von ihren Verwendungen lehren!
Laß dich die Bedeutung durch den Gebrauch lehren
Ludwig Wittgenstein

You shall know a word by the company it keeps
John R. Firth

[…] the basic units are occurrences of the word in context
Adam Kilgarriff

Resumen

El concepto de *perfil combinatorio* (*behavioral profile, lexical profile, perfil comportamental, profil combinatoire*) es central en las investigaciones ligadas a trabajo empírico con elementos lingüísticos, esencialmente llevadas a cabo a partir de los datos extraídos de corpus. En este capítulo se revisan las definiciones de *perfil combinatorio* que se encuentran en la bibliografía —desde Hanks (1996)— y se plantea una reflexión sobre la importancia del desarrollo del concepto, así como sobre la relevancia de otorgar mayor peso a la combinatoria léxica sobre la gramatical en los estudios de carácter contrastivo.

Palabras clave: perfil combinatorio, esquema sintáctico, polisemia, sinonimia, lexicografía, lingüística de corpus

Abstract

The concept of *behavioral profile* (*lexical profile, perfil combinatorio, perfil comportamental, profil combinatoire*) is central to research concerning empirical work

with linguistic elements, typically carried out with data extracted from corpora. This chapter reviews the definitions of *behavioral profile* found in the bibliography —since Hanks (1996)—, reflects on the importance of the further development of the concept and considers the relevance of attaching greater weight to lexical than to grammatical combinations in contrastive studies.

Keywords: behavioral profile, syntactic pattern, polysemy, synonymy, lexicography, corpus linguistics

1. Introducción

La posibilidad de acceder a gran cantidad de datos lingüísticos de manera automática, en lo que constituyó, a principios de los años 90 del siglo XX, el desarrollo de la lingüística basada en el análisis de corpus, supuso el entorno propicio para el nacimiento del concepto de *perfil combinatorio*[1]. En los inicios de la lingüística de corpus, la explotación de corpus textuales tuvo como elemento central el recurso de la obtención de concordancias. Conforme las colecciones de textos de tipología y características diversas se hacían mayores, tanto la extracción de fragmentos destinados a determinar el significado de una palabra —un grupo de palabras, un elemento gramatical— o a ejemplificar su uso, como el recuento de formas o estructuras determinadas para aportar datos cuantitativos se convirtieron en la base de los análisis lingüísticos de corte empírico (Rojo 2015: 687). La obtención de concordancias y el recuento de elementos lingüísticos son los dos procedimientos que soportan en su fundamento la idea de perfil combinatorio.

Por otra parte, se trata de un concepto surgido en el ámbito de la lexicografía, de manera prioritaria desde una perspectiva semasiológica. Por ello, veremos que uno de los aspectos clave del concepto, y de las metodologías en las que se recurre a él, es la relación que se establece entre los significados o acepciones de un lema y las diversas construcciones sintácticas de las que participa. El recurso a las definiciones lexicográficas en conexión con el análisis refinado de las concordancias extraídas de corpus es lugar común en los estudios empíricos, aunque la cuestión de apelar a la introspección de quien lleva a cabo la investigación, así como la adopción de una perspec-

[1] Este capítulo se ha llevado a cabo en el marco del proyecto COMBIDIGILEX («La combinatoria en paradigmas léxico-semánticos en contraste. Estudio empírico y digitalización para el aprendizaje de lenguas extranjeras en el contexto germano-iberorrománico»). El proyecto ha contado con la siguiente financiación entre los años 2016-2019: FEDER - Ministerio de Ciencia, Innovación y Universidad - Agencia estatal de Investigación FFI2015-64476-P / COMBIDIGILEX.

tiva de síntesis no se desatienden cuando el punto de vista es de carácter onomasiológico.

En las casi dos décadas transcurridas del siglo actual, este concepto se ha convertido en un eje metodológico, pues constituye el ámbito de diferentes estudios inter- e intralingüísticos de carácter empírico, cuyo objetivo es precisamente el establecimiento de (algunas de) las coordenadas combinatorias de unidades lingüísticas de muy diversa naturaleza, atendiendo también, en ocasiones con preferencia, a los datos cuantitativos relativos a la frecuencia y a análisis estadísticos más o menos sofisticados. A modo de ejemplo, podemos mencionar algunos estudios recientes que analizan unidades como las siguientes: lexemas concretos, como el verbo *sentir* (Jansegers 2017), unidades multipalabra, como el fraseologismo *por momentos* (Mellado & López Meirama 2017), clases semánticas verbales, como la clase «verbos de competición», manifestada en cuatro de sus representantes —*ganar, vencer, derrotar* y *perder*— (García-Miguel 2014), o un determinado elemento de la conjugación verbal, como las formas verbales del subjuntivo en *-ra* en el español de Galicia (Rojo & Vázquez Rozas 2014). Todos estos trabajos se sirven de datos obtenidos de corpus textuales diversos, otorgando especial relevancia al comportamiento de las unidades lingüísticas dentro de contextos textuales reales y a la frecuencia de las coocurrencias léxicas, gramaticales o discursivas, con objeto de analizar y descifrar los significados de las unidades estudiadas en cuanto a fenómenos de polisemia, sinonimia, variación y cambio lingüístico.

La estructura de este capítulo es la siguiente: tomando como punto de partida las definiciones iniciales de perfil combinatorio y una selección de ejemplos que se encuentran en la bibliografía (2), atendemos a continuación a algunas aportaciones destacadas de la aplicación metodológica del concepto (3), para concluir con las reflexiones derivadas del recorrido trazado y considerar, finalmente, los principales retos que se plantean en lo relativo a las aproximaciones de carácter contrastivo (4).

2. Qué se entiende por *perfil combinatorio*

2.1. La preocupación por obtener la posición integral de un elemento lingüístico de una lengua en concreto —un elemento léxico o gramatical, simple o compuesto— determinó desde el primer momento la acuñación del concepto *behavioral profile*. Según las reflexiones de Hanks (1996: 78), los datos extraídos de corpus textuales permiten establecer los que se podrían considerar los esquemas de uso normal, central y típico —e incluso permiten, decía, con la

debida precaución, precisar cuáles son los empleos más frecuentes—; pero el autor insiste en el hecho de que no es posible obtener la deseada posición integral, esto es, una relación de todos los usos posibles de un ítem, ya que, afirma, ninguna cantidad de corpus puede probar que algún fenómeno lingüístico no pueda existir. He aquí una primera limitación del trabajo empírico: cuando hablamos de perfil combinatorio, nos referimos entonces a *un perfil combinatorio*, según las características del corpus o los corpus tenidos en cuenta.

La argumentación y discusión presentada por Hanks (1996) se ilustra con el recurso a algunos ejemplos de perfiles combinatorios, como el que se reproduce en la figura 1.

Lemma: **incite, incites, inciting, incited**
One pattern with three variations:

0.1. [PERSON *or* SOMETHING] incites [PERSON] to-INF [DO [BAD]]
0.2. [PERSON *or* SOMETHING] incites [PERSON] to-PREP [ACTION *or* ATTITUDE [BAD]]
0.3. [PERSON *or* SOMETHING] incites [ACTION or ATTITUDE [BAD]]

lexical items in the set [DO [BAD]] found in this context include: *rebel, revolt, go on strike, assassinate, be naughty, break the law, commit [CRIME], go shoplifting, [VP] illegally, breach [NP]*

lexical items in the set [ACTION [BAD]] found in this context include: *riot, arson, debauchery, discord, dissention, denunciation, hatred, crime, lewdness, murder, trouble, unrest, violence, revolution, demonstration.*

FIGURA 1. Perfil combinatorio del verbo inglés *incite* (Hanks 1996: 87)

El perfil combinatorio de *incite* es la suma de los patrones de complementación de ese lema, en los que se indican las características semánticas generales de los actantes —PERSON, SOMETHING, ACTION, ATTITUDE, BAD—, algunas características gramaticales, como el tipo de unidad o la manifestación sintáctica —to-INF, to-PREP—, y una relación de ejemplos de elementos léxicos concretos, uni- o pluriverbales, que desempeñan determinadas funciones, obtenidos como resultado de una búsqueda en el *British National Corpus* (BNC). En este ejemplo con *incite* no se indican los datos cuantitativos —de frecuencia—, aunque sí se tendrán en cuenta, pues el número de apariciones del verbo en el corpus con cada uno de los patrones se considera una información relevante.

Otro de los ejemplos que se ofrecen en el artículo citado es el perfil combinatorio del verbo inglés *urge* —según los datos extraídos del Hector Pro-

ject—, lo que, en palabras de Hanks, constituye: «an attempt to encapsulate its established norms (patterns of usage) on the basis of analysis of a body of evidence of actual usage (a corpus)» (Hanks 1996: 79). Se trata entonces de mostrar de manera esquemática o resumida los patrones sintácticos y de coocurrencia léxica en los que la palabra en cuestión se manifiesta, siempre según la evidencia obtenida a partir de las concordancias de un corpus determinado, que brindan la posibilidad de detectar cuáles son los empleos más frecuentes —más comunes. Los patrones se ordenan de más a menos convencionales, teniendo siempre presentes dos factores: primero, no se trata de *todos* los patrones posibles, puesto que, como ya ha quedado dicho, no es factible tomar en consideración todas las manifestaciones lingüísticas, y, segundo, los datos de frecuencia presentan siempre un desequilibrio entre unos pocos esquemas muy frecuentes y varios esquemas más escasamente comunes (según formula la ley de Zipf).

Nótese que en la figura 1, al igual que en los otros ejemplos de Hanks (1996), lo que no se muestra de manera explícita en el perfil es propiamente una definición ni explicación de cada uno de estos patrones y variantes, como tampoco se ofrece ningún tipo de equivalencia sintáctico-semántica para el lema, de manera que, como el mismo autor subraya, en el perfil combinatorio no se dice nada sobre significados, definiciones ni traducciones del ítem léxico. Por supuesto que, en su artículo, Hanks ya ponía el énfasis en la relación existente entre la semántica de una unidad léxica y la totalidad de sus patrones de complementación, es decir, entre los significados de un lexema —él se refiere a un verbo— y las construcciones en las que participa. De hecho, el autor parte del presupuesto de que el potencial de significado de cada verbo viene determinado, al menos en parte, por la totalidad de los contextos en los que aparece (Hanks 1996: 90). Así pues, se postula la premisa metodológica de crear un diccionario sin definiciones, consistente en que para cada lema se encuentra un reducido número de patrones o esquemas, que vale por los usos convencionales y constituye una proporción muy elevada —entre el 70 y el 80% o más— de todos los usos documentados (Hanks 1996: 84). A partir de ahí, la labor lexicográfica se nutre de los datos empíricos, estableciéndose un recorrido circular entre el diccionario y el corpus, siendo el segundo una parte cada vez menos prescindible, tanto si pensamos en una orientación que pretenda poner el diccionario en el corpus como en la que se plantee poner el corpus en el diccionario (Kilgarriff 2005, Alonso-Ramos 2009).

2.2. Durante la última década del siglo XX tomó forma la *Base de datos sintácticos del español actual* (BDS en adelante), el resultado de la codificación del

análisis sintáctico manual minucioso de aproximadamente 160.000 cláusulas (de un corpus de un millón y medio de palabras). Se trata de un proyecto que, aunque no se refería expresamente al concepto de perfil combinatorio, se proponía ofrecer, en una de sus posibles explotaciones, una relación de los esquemas y subesquemas en que aparece cada uno de los verbos contenidos en el corpus de partida (Rojo 2001, Mas & Rojo 2004). En la figura 2 se muestra la pantalla de respuesta de una búsqueda en la BDS de los esquemas del verbo *sentir*. Los datos se corresponden, como es sencillo deducir, con las siguientes características: voz (activa/media/pasiva), funciones sintácticas de carácter argumental (regidas o valenciales), número total de ejemplos del corpus correspondientes al verbo *sentir* en ese esquema y frecuencia (porcentaje con respecto al total del verbo)[2]. Por supuesto, la BDS contempla además el acceso a los ejemplos —las concordancias—, de manera que es sencillo visualizar los verbos en su contexto y desplazarse así de la presentación más abstracta de cada esquema o subesquema a la concreción de las secuencias reales recogidas en los textos que componen el corpus.

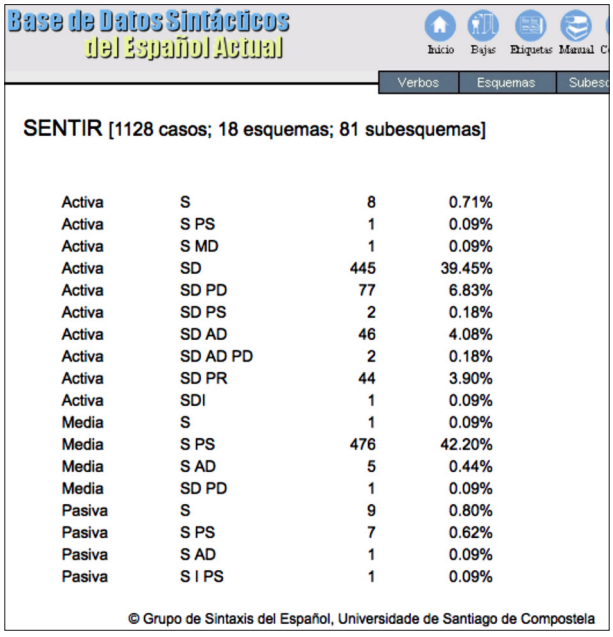


FIGURA 2. Esquemas en los que entra el verbo *sentir* (BDS)

[2] Clave de las abreviaciones de las funciones sintácticas: AD = complemento adverbial regido, D = complemento directo, I = complemento indirecto, MD = complemento modal regido, PD = complemento predicativo del complemento directo, PR = complemento preposicional regido, PS = complemento predicativo del sujeto, S = sujeto.

La información correspondiente a los subesquemas incluye datos más detallados, como el tipo y subtipo de unidad que desempeña cada función sintáctica —frases, cláusulas, o palabras gramaticales de diversos tipos—, así como el carácter animado o inanimado del elemento que desempeña cada función. El despliegue en subesquemas mantiene, por lo tanto, un paralelismo con las variantes de los esquemas diferenciadas por Hanks (1996), aunque aquí no se agrupan bajo una etiqueta semántica. Como se puede apreciar en la figura 2, los valores frecuenciales cumplen el reparto habitual que el mismo Hanks señalaba, pues encontramos que las frecuencias de uso se acumulan en unos pocos esquemas, con una gran distancia respecto al resto de los empleos. En el caso de este verbo concreto, podemos establecer una escala de tres estadios de frecuencia, ya que algunos esquemas de frecuencia intermedia pueden considerarse extensiones de los más frecuentes: el esquema activo más frecuente, SD, presenta tres extensiones a base de elementos sintácticos diferentes, SD PD, SD AD y SD PR.

Pero el diseño de la BDS permite otras explotaciones también interesantes, porque anticipan una concepción más amplia del perfil combinatorio: en lugar de partir de un verbo determinado para las búsquedas, es posible partir de un esquema o de un subesquema para tener como resultado un listado de los verbos que aparecen en ese esquema o subesquema, con los datos de frecuencia. En este caso vamos de la construcción a los verbos concretos que desempeñan el papel de predicado en esa construcción, es decir, conseguimos un catálogo de los verbos que, en un corpus determinado, pueden aparecer en un mismo contexto sintáctico. Ello permite considerar el perfil combinatorio de una construcción, para el que sería deseable reunir, quizá en diferentes niveles de generalización, qué elementos cubren cada uno de los lugares funcionales. Sin embargo, la BDS no proporciona los elementos léxicos concretos que ocupan cada una de las funciones sintácticas, por lo que para detectar la combinatoria léxica, más o menos preferida, con un determinado verbo es necesario revisar los ejemplos uno a uno dentro de cada subesquema.

Por último, la BDS no contiene información semántica. El camino hacia un diccionario estaba iniciado, pero la tarea se vio como un paso ulterior, puesto que el objetivo principal lo constituyó siempre el análisis de la información sintáctica a partir del verbo: «es claro que la visión real del comportamiento sintáctico de un verbo aparece cuando somos capaces de poner en relación los cambios en el esquema o subesquema sintáctico utilizado con las diferencias en el significado» (Rojo 2001: 281). La pretensión era vincular los ejemplos de cada verbo a las acepciones y subacepciones, de manera

que fuera factible presentar la doble organización de la información, a la que se llegaría bien partiendo de la construcción sintáctica, bien del significado, mostrando así claramente la relación existente entre las acepciones y los subesquemas sintácticos, que se complementarían con los datos frecuenciales. El desarrollo de la BDS tomando en consideración información de tipo semántico tuvo lugar en los años sucesivos a través del proyecto ADESSE, sobre el que volveremos enseguida.

2.3. El empleo del concepto de perfil combinatorio cuenta con una vía de investigación relativa a la diferenciación de sinónimos desde el trabajo de Blumenthal (2002). El autor aporta la siguiente definición de *profil combinatoire*: «structure schématique du voisinage syntaxique et sémantique d'un mot telle qu'elle se manifeste dans un vaste corpus» (Blumenthal 2002: 115-116). En esencia, los presupuestos de partida aludidos en las concepciones anteriores están contemplados aquí también: el carácter abstracto o general de la estructura —esquemático—, la coocurrencia de elementos como premisa esencial —vecindad, proximidad— y la manifestación en un corpus amplio[3]. A estos presupuestos hay que sumar en la definición de Blumenthal la mención explícita al hecho de que se trata de una estructura tanto sintáctica como semántica.

La finalidad de la aproximación de este autor es doble: por una parte, le interesa indagar en las diferencias que hay entre sustantivos sinónimos de una misma lengua tomando como base un estudio detallado de sus contextos; por otra, se propone profundizar en la caracterización de los sentidos de un sustantivo de una lengua según sus supuestos sustantivos equivalentes en otra, de manera que se puedan establecer los contrastes interlingüísticos pertinentes (Blumenthal 2002: 115). El terreno concreto es el estudio de la sinonimia entre sustantivos del campo semántico del alemán *Angst* y el francés *peur* y el punto de partida, tras las precisiones terminológicas y conceptuales, es el diccionario. La identificación del perfil combinatorio, concluye el autor, resulta de las valencias activa y pasiva de los sustantivos, de sus relaciones sintagmáticas no valenciales así como de los escenarios en los que aparecen. Su interpretación ofrece la posibilidad de definir un perfil aspectual, un perfil ontológico, un perfil paradigmático y un perfil de satu-

[3] La distancia entre lo que se consideraba un corpus amplio a principios de este siglo y lo que se considera un corpus representativo en cantidad en la actualidad es notable, desde luego. Paradójicamente, cuanto más amplios son los corpus de referencia empleados en los análisis lingüísticos empíricos más necesario se hace reducirlos, dada la inevitable parte manual de análisis y revisión que todavía requiere esta metodología de trabajo.

ración, los cuales representan parámetros semánticos parcialmente medibles (Blumenthal 2002: 136).

La concepción de Blumenthal, que él mismo continúa desarrollando en trabajos posteriores, es la referencia de algunas publicaciones dedicadas al estudio de la combinatoria léxica de verbos, sustantivos y adjetivos, parte de ellas interesadas por determinados campos semánticos (como el de la emoción), por la sinonimia y por el análisis contrastivo (Parte I, dedicada a la combinatoria léxica, de Bailer & Cislaru 2013). Se trata igualmente de aproximaciones basadas en corpus, incluyendo corpus paralelos en la perspectiva contrastiva, que comparan los perfiles combinatorios de palabras concretas, recurriendo preferentemente a las estructuras actanciales y a la coocurrencia sintáctico-semántica (ocasionalmente también a la polaridad).

2.4. Las posturas resumidas en los apartados 2.1. y 2.2. han conocido desarrollos en los que la dimensión semántica ha entrado de lleno. En Hanks (2013) se presenta la *Theory of Norms and Exploitations* (TNE), que parte de un núcleo de concordancias seleccionadas de corpus y otros textos para mostrar cómo es posible emparejar usos y patrones contextuales para determinar el significado de las secuencias. Este lingüista ha desarrollado la metodología llamada *Corpus Pattern Analysis* (CPA) en el seno de un proyecto cuya finalidad primordial es explorar la relación entre el significado de las palabras y sus patrones de uso, es decir, las palabras en contexto. Se pretende identificar todos los perfiles de uso normal de los verbos ingleses a través de un análisis empírico sistemático del BNC —y distinguir los usos normales, convencionales (*norms*), de los originales y creativos (*exploitations*)—, pues hay la certeza de que los significados verbales se corresponden con patrones fraseológicos y no con listas de verbos aislados (Hanks 2013: 404-405). Es posible consultar en línea el resultado de este trabajo, que se encuentra en proceso de desarrollo (PDEV). En la figura 3 se presenta el perfil del verbo *incite*. El acceso completo a los datos enlaza con las 152 concordancias del BNC, para cada una de las cuales se indica el número de patrón con el que se corresponde. El enlace a más datos muestra los ejemplos de cada patrón concreto. Nótese que este diccionario proporciona información también de la frecuencia relativa de cada esquema. Los actantes de cada patrón están numerados y se emplean para ellos etiquetas generales —*Human, Institution, Action=Bad*—, las cuales agrupan los elementos concretos de la combinatoria por su valor semántico. En el proceso de elaboración de las entradas del diccionario, los contextos en los que aparece el lema dentro del corpus —las concordancias— se agrupan

en patrones sintagmáticos semánticamente motivados; en un segundo paso se asocia un significado con cada patrón.

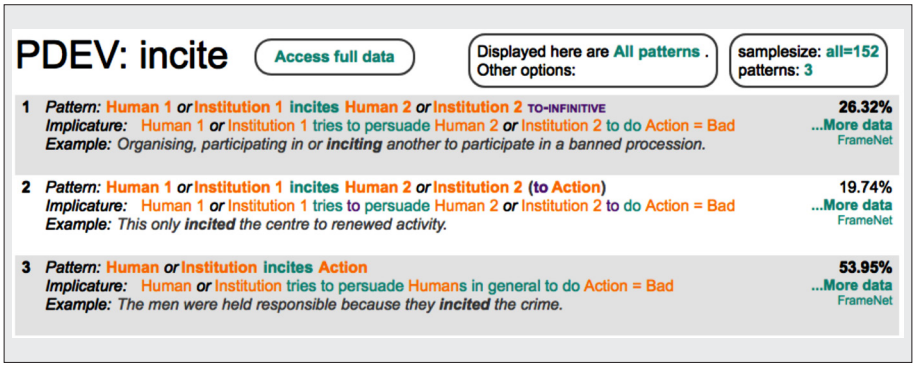


FIGURA 3. Entrada del verbo *incite* (PDEV)

Por lo que se refiere a la BDS, ya apuntábamos más arriba que encontró una ampliación en el proyecto de una *Base de datos de Verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español*, accesible en línea a través de un sistema de búsquedas enriquecido. Con el mismo corpus y la información sintáctica ya elaborada, ADESSE incorpora la información semántica en tres niveles: agrega los papeles semánticos de los actantes —con una numeración correlativa y etiquetas—, asigna los ejemplos a las acepciones de los verbos en diferentes niveles de generalidad y clasifica las acepciones de los verbos en una tipología de procesos o clases semánticas (García-Miguel & Albertuz 2005). De entre las posibilidades de acceso a los datos de ADESSE interesa aquí la que ofrece un perfil combinatorio, sea para cada acepción de un verbo, sea para cada esquema sintáctico-semántico.

El perfil combinatorio del verbo en una acepción determinada está encabezado por una definición más la indicación del número total de ejemplos anotados del corpus. A continuación, se presenta la información del perfil desplegada en un resumen cuantitativo de las propiedades de los argumentos. En la tabla 1 se reproduce el perfil combinatorio de SENTIR II - *Lamentar [algo] (que ha ocurrido)*, una acepción que cuenta con 69 ejemplos anotados en el corpus.

Argumento: *(Frec. explícito)*	**A1 (Experimentador) 69 (100%)**		**A2 (Estímulo) 69 100%**		**A3 (Causa) 1 (1%)**	
Función sintáctica:	**SUJ**	69	**ODIR**	69	**OBL**	
Clíticos objeto:			**ACUS**	55		
Preposiciones:	ø	4	ø	14	por	1
Categoría sintáctica:	FN Pro	2 2	Claus-inf FN Claus-subj	8 4 2	Pro	1
Tipo semántico:	**Animado**	69	**Proposicional** Abstracto Concreto	59 9 1	Animado	1
Realizaciones léxicas frecuentes: *(solo no animados)*			ø *decir* *ir* *entretener* *mandar* *coger* *empezar* *herir* *tener*	58 2 2 1 1 1 1 1 1		

Tabla 1. Perfil combinatorio de SENTIR II (ADESSE)

Este cuadro resumen muestra de manera sinóptica la información sintáctico-semántica en distintos niveles de generalización y, además, incluye información léxica sobre los elementos nucleares concretos de la combinatoria. En este ejemplo, los predicados de las cláusulas de infinitivo o en subjuntivo en función de objeto. Clicando en el dato numérico se accede a las secuencias concretas del corpus que cumplen la propiedad sintáctico-semántica y, en aquellos casos en que exista, la realización léxica concreta recogida en el resumen.

Para García-Miguel, el perfil combinatorio es «el conjunto de probabilidades de coocurrencia con elementos léxicos y gramaticales de cualquier nivel de generalidad que presenta una unidad lingüística cualquiera» (García-Miguel 2014: 14). Este conjunto de probabilidades se presenta como cerrado, pues se extrae del análisis y categorización de datos desde un corpus también cerrado; sin embargo, debe entenderse como dinámico, lo mismo que debe entenderse como dinámico el uso que de los elementos lingüísticos —con sus combinaciones— hacemos las personas, además de que es también dinámico el significado.

Más arriba nos hemos referido a la circularidad implicada en la labor lexicográfica, que va del diccionario al corpus y del corpus otra vez al diccio-

nario, o viceversa. Parece que el trabajo empírico obliga a este mismo procedimiento, que debe tener en cuenta tanto el «significado potencial» de la unidad estudiada, como el de las estructuras sintácticas en las que puede combinarse, es decir, las construcciones, sin olvidar los elementos léxicos concretos preferidos para ocupar los lugares funcionales o las condiciones de uso. Cuál debe ser el punto de partida no resulta claro, pues en general viene determinado por la concepción adoptada, semasiológica u onomasiológica, y por otros factores relacionados con la finalidad de cada análisis: cuáles son las unidades estudiadas y de qué clase —verbos frente a sustantivos, fraseologismos de diverso tipo, por ejemplo—, si se trata de un análisis contrastivo —y en este caso la distancia tipológica entre las lenguas implicadas es decisiva—, si el objetivo es indagar sobre un fenómeno de variación, etc.

Uno de los factores relevantes en la concepción del perfil combinatorio de una unidad lingüística es, desde luego, la coocurrencia de elementos léxicos. El tamaño del corpus y las características de los textos que lo componen condicionan en gran medida la cantidad, variedad y cualidad de las unidades léxicas concretas que podamos obtener para un determinado hueco funcional en las concordancias, pero el agrupamiento de esos elementos por su significado permite un grado de generalización que simplifica el perfil y lo convierte en más representativo de su valor potencial. A ello se refiere García-Miguel a propósito de la selección argumental cuando ilustra su explicación a partir de los datos que ofrece ADESSE sobre el verbo *vencer*: «[…] lo que resulta más significativo no es la combinatoria con lexemas individuales sino con conjuntos de lexemas semánticamente relacionados. Lo que se vence no son solo obstáculos, sino también dificultades, rémoras, crisis y otras situaciones negativas.» (García-Miguel 2014: 33).

2.5. En la definición de perfil combinatorio es necesario entonces tener en cuenta al menos los elementos siguientes: corpus, concordancias, coocurrencia de rasgos sintácticos, semánticos, léxicos y discursivos en distintos niveles de generalización, y frecuencia. La referencia a uno o varios corpus garantiza que el punto de partida es el uso real de los elementos lingüísticos; el valor del corpus es, por tanto, el de constituir una evidencia de tal uso: la representatividad y características del corpus condicionarán la representatividad y características del perfil combinatorio, por lo que la elección del corpus o los corpus debe estar en consonancia con los objetivos de cada investigación. Por otra parte, la ayuda de la tecnología para automatizar cometidos como la compilación, procesamiento y alineamiento de corpus permite que la creación de corpus *ad hoc* resulte cada vez más accesible.

Las concordancias son la vía de acceso a las unidades concretas a partir de las cuales se procederá a la categorización y agrupación de elementos, con la palabra o palabras clave como eje. Para facilitar en lo posible la tarea del análisis empírico es deseable contar con una etiquetación y desambiguación del corpus de calidad, ya que en la medida en que se presenten más o menos deficiencias la tarea resultará más o menos ágil. Aún queda mucho por mejorar, dependiendo de las lenguas con las que se trabaje. La calidad de la anotación del corpus y el grado de detalle inciden en los datos registrados en las coocurrencias, que pueden ser de muy diverso tipo. Los parámetros morfosintácticos, semánticos, léxicos y pragmático-discursivos pueden requerir un grado de detalle para el que el análisis automático no está lo suficientemente desarrollado. Por último, la frecuencia es la base de las técnicas de estadística distribucional cada vez más refinadas que se aplican a unidades simples o compuestas, con mayor o menor grado de abstracción. Veamos en el apartado siguiente cómo se está optimizando el rendimiento del recurso al perfil combinatorio.

3. El perfil combinatorio como metodología

A lo largo de las primeras décadas de este siglo el concepto de perfil combinatorio se ha ido enriqueciendo y ha llegado a convertirse en el eje de metodologías concretas en el ámbito de los estudios basados en corpus. De una parte se han desarrollado programas de búsqueda y de análisis estadístico que resultan de gran ayuda para el trabajo empírico, ya que han desarrollado técnicas diversas para medir estadísticamente las coocurrencias significativas; de otra, los objetivos de los estudios se han ido diversificando, demostrando que una metodología que tiene como eje la elaboración y análisis de perfiles combinatorios resulta fructífera para describir hechos lingüísticos. Los diversos acercamientos tienen como objetivo aplicaciones de diferente tipo, siendo las investigaciones contrastivas y sobre fraseologismos algunas de las que más se benefician de esta perspectiva, pues presentan gran utilidad en los ámbitos de la didáctica de lenguas, la traducción o la lexicografía. Entre los primeros recursos para explotar corpus textuales aplicando refinadas mediciones estadísticas destaca *Sketch Engine*, al que se dedican los párrafos siguientes. Entre los estudios, resumiremos después las características de los acercamientos llevados a cabo con el método del perfil combinatorio.

3.1. La herramienta *Sketch Engine* fue lanzada en 2004, pensada en principio como ayuda para el trabajo lexicográfico. Tras más de diez años, caracteriza-

dos por una auténtica revolución en el terreno de la lexicografía y un avance en el diseño y explotación de corpus textuales, sigue afrontando el reto de adecuar sus funciones a las necesidades de quienes se interesan profesionalmente por las lenguas desde ámbitos diferentes: la traducción, la lexicografía, la lingüística sincrónica y diacrónica, la terminología, la didáctica de lenguas, la adquisición de lenguas, la tecnología lingüística (Kilgarriff *et al.* 2014)[4]. Como es sabido, *Sketch Engine* ofrece un conjunto de recursos informáticos destinados al análisis estadístico del lenguaje en uso, para identificar aquello que es frecuente y típico en una lengua determinada. Mediante la búsqueda de lemas en corpus textuales de miles de millones de palabras, se obtienen las combinaciones, tanto léxicas como gramaticales, más típicas del lema, junto a las menos frecuentes, las raras o las que insinúan un uso emergente. Se trata de un sofisticado conjunto de recursos que tiene la virtud de parecer sencillo y, sobre todo, que es capaz de devolver instantáneamente el resultado de las búsquedas, ofreciendo opciones de visualización novedosas e interesantes. También las posibilidades de solicitud de datos son variadas, siempre tomando como base las palabras en su contexto. Dentro del variado conjunto de posibilidades que brinda el recurso, me voy a referir tan solo a las tres opciones que conectan más directamente con la idea de perfil combinatorio: *Word Sketch, Word Sketch Difference* y *Thesaurus*.

Lo que se ofrece a través de la opción *Word Sketch* es un resumen del comportamiento de la unidad buscada —simple o multipalabra—, de manera que se obtiene la información pertinente que nos permite indagar en un perfil combinatorio y analizarlo. En la visualización del resumen es posible ordenar los resultados, si se desea, por la frecuencia o por índice de tipicidad (*score*)[5], dos valores que pueden resultar discordantes. Dependiendo de cuál sea la unidad requerida, el resultado del esquema aporta distintas categorías. Ilustremos la funcionalidad de *Word Sketch* desde algunos de los resultados del resumen del lema «sentir», que muestran el rico perfil combinatorio esperable en un verbo polisémico de uso tan frecuente en español, con categorías como las siguientes: seguido de adjetivo, seguido de

[4] El acceso a la página web de *Sketch Engine* y a todas sus funciones es gratuito para instituciones académicas desde abril de 2018 hasta abril de 2022 —siempre para uso no comercial— gracias a la financiación de ELEXIS (*European Lexicographic Infrastructure*), organismo que cuenta con una ayuda del programa de investigación e innovación *Horizonte 2020* de la Unión Europea.

[5] El índice de tipicidad —*score*— mide el grado de fijación de la coocurrencia y responde a análisis estadísticos de diverso tipo: cuanto más alta es la puntuación, más estrecha es la relación entre las unidades que coocurren, que cuenta como más típica; una puntuación baja indica que las unidades que coocurren se combinan también frecuentemente con muchas otras palabras, por lo que su relación léxica es menos estrecha o no tan preferida, es decir, se trata de una combinación no típica.

adverbio modificador, seguido de sustantivo —objeto directo—, sujetos de «sentir», «sentir» y/o..., frases preposicionales combinadas con «sentir», pronombres objeto, pronombres sujeto, *wh-words* que siguen a «sentir», verbos en infinitivo tras «sentir», verbos en gerundio tras «sentir», preposiciones que siguen a «sentir», y patrones gramaticales de uso, como indicativo/subjuntivo, singular/plural, reflexivo/pasivo, infinitivo/gerundio/participio, en perífrasis, etc. En la figura 4 tenemos tres ejemplos de los elementos léxicos concretos que se combinan más típicamente con el lema en tres patrones sintácticos diferentes. Conocemos de manera instantánea los adjetivos, adverbios y sustantivos con los que se coloca *sentir*, con indicación del número de ocurrencias que presentan el corpus y la puntuación de tipicidad. Apreciamos aquí cómo el índice de tipicidad no es paralelo al de frecuencia: el adjetivo *solo*, el adverbio *muy* o el sustantivo *amor*, por ejemplo, presentan sí una frecuencia de aparición elevada en combinación con *sentir*, pero, en este corpus al menos, también se combinan frecuentemente con otros verbos, por lo que no son tan «preferidos», por ejemplo, como *libre*, *profundamente* o *vergüenza*, respectivamente.

adjectives after "sentir"				modifiers of "sentir"				objects of "sentir"			
cómodo	754	11.27	•••	mal	591	9.34	•••	dolor	467	9.58	•••
orgulloso	627	11.13	•••	mejor	335	8.44	•••	necesidad	440	8.9	•••
culpable	262	10.03	•••	bien	1,023	8.03	•••	miedo	256	8.63	•••
incómodo	200	9.76	•••	seguro	90	7.58	•••	vergüenza	129	8.13	•••
feliz	291	9.48	•••	profundamente	72	7.37	•••	placer	139	8.1	•••
seguro	286	9.45	•••	más	1,497	7.12	•••	emoción	139	8.08	•••
libre	198	9.06	•••	realmente	145	7.07	•••	curiosidad	124	8.05	•••
inseguro	105	8.93	•••	cerca	95	7.04	•••	calor	126	7.98	•••
triste	126	8.92	•••	muy	1,228	6.73	•••	sensación	153	7.94	•••
solo	230	8.72	•••	tanto	471	6.44	•••	atracción	100	7.81	•••
mejor	221	8.49	•••	igual	50	6.41	•••	orgulloso	95	7.79	•••
especial	88	8.39	•••	así	224	6.39	•••	amor	130	7.67	•••

FIGURA 4. Algunos de los elementos destacados de la combinatoria léxica del verbo sentir (*Spanish Web 2018 sample en Sketch Engine*)[6]

Como ya se ha apuntado más arriba, el trabajo empírico requiere revisar con cierto detalle las concordancias de cada uno de los resultados que arrojan las búsquedas, pues, desafortunadamente, con frecuencia el análisis auto-

[6] El corpus sobre el que se ha lanzado esta búsqueda es una muestra del corpus *Spanish Web 2018* (177257648 palabras: 1/100 sample of the Spanish web corpus crawled by Spideling in February to April 2018. Tokenised, lemmatised. Tagged by Freeling pipeline v. 5. Text sources: Spanish Wikipedia, European Spanish web, American Spanish web).

mático esconde errores de diverso tipo, a veces debidos a que los módulos de lematización y etiquetado son aún mejorables; en otras ocasiones por duplicaciones o desajustes en la lematización asociados a una grafía inesperada. En lo que se refiere a *Sketch Engine*, los problemas con el español afectan a varios aspectos trascendentales, algunos de los cuales se pueden descubrir desde los datos de la figura 4. El más llamativo es el hecho de que el lema «orgulloso» aparece listado en una relación de sustantivos objeto de «sentir». Al consultar las concordancias se pone de manifiesto que la asignación de la función de objeto se hace automáticamente en cláusulas de infinitivo sobre la forma que sigue al infinitivo, aunque no se trate de un sustantivo. Por supuesto, el lema «orgullo» también está en la relación de sustantivos objeto, con 84 apariciones. Por otra parte, el sincretismo existente en ciertas formas de la conjugación de «sentir» y «sentar» —*siento, sientes, siente, sienten, sienta, sientas, sientan*—, junto con la coincidencia de algunos de sus esquemas sintácticos y colocaciones —sentir(se)/sentar(se)/sentar + bien/mal/mejor— hace que se fusionen los ejemplos, lo cual enturbia los resultados y resta utilidad al potencial de la herramienta.

A partir de los resultados de la figura 4 es factible hacer un zum hacia el fenómeno combinatorio concreto que sea de nuestro interés, como por ejemplo el *Word Sketch* de la colocación «sentir necesidad» (figura 5). Los modificadores de «sentir necesidad» se desglosan en los que afectan a cada uno de los elementos de la colocación.

modifiers of "sentir ... necesidad"							
"sentir"				"necesidad"			
nunca	6	2,37	•••	imperioso	8	9,2	•••
entonces	3	1,82	•••	urgente	3	4,54	•••
más	12	0,2	•••	fuerte	3	2,13	•••

FIGURA 5. *Word Sketch* de «sentir necesidad» (*Spanish Web 2018 sample en Sketch Engine*)

El recurso *Word Sketch Difference* es una interesante opción cuando buscamos la comparación entre dos lemas. En el ejemplo de la figura 6, mostramos algunos datos de la comparación entre dos verbos muy próximos en sus perfiles sintáctico-semánticos y cercanos también en cuanto a su significado: los verbos *quitar* y *sacar*. En español estos verbos expresan 'desposesión' y 'desplazamiento', por lo que coocurren con otros verbos de semántica similar, listados en la columna «quitar/sacar and/or...». Esta

lista refleja un reparto claro de las coocurrencias, que solo comparten el verbo *poner*, aunque con clara diferencia en frecuencia: se aprecia que *quitar* se combina sobre todo con *poner* (y *colocar*), aunque también con *dar* (*agregar, añadir*), y que *sacar* se combina principalmente con *meter*. Estos verbos presentan además gran frecuencia de uso como verbos soporte en colocaciones con sustantivos objeto. Los sustantivos con los que se colocan se ofrecen en la comparación de los «objects of quitar/sacar», que refleja el valor de *sacar*, siempre con más ejemplos, como base de construcciones con verbos de apoyo (Mas Álvarez 2018).

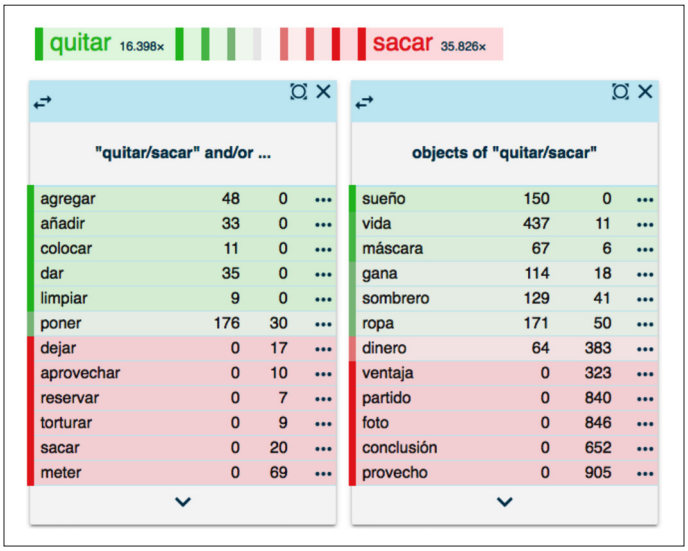


FIGURA 6. *Word Sketch* Difference resumido de los lemas quitar y sacar (*Spanish Web 2018 sample en Sketch Engine*)

Como vemos, esta opción de recuperación de los datos es idónea para identificar solapamientos de perfiles, que podrían apuntar a una sinonimia parcial o, teniendo en cuenta la procedencia de los textos, a un fenómeno de variación. Como el único verbo compartido, en este corpus concreto, en las combinaciones de *quitar* y *sacar* es *poner*, que coaparece mucho más frecuentemente con *quitar*, el acceso a los ejemplos avala la nota diferencial que incorpora *sacar* en el sentido del desplazamiento, pues se refiere a los casos en que el desplazamiento se produce ‘de dentro hacia fuera’: sacar y poner a un jugador, una rueda, un accesorio, un motor, un techo (en un vehículo), partes del cuerpo, un cartucho en la consola, por ejemplo.

La proximidad sintáctico-semántica de los verbos *quitar* y *sacar* ha servido como explicación de sus usos intercambiados en determinados dialectos del español hablado a ambos lados del Atlántico, el español de Galicia entre ellos (Mas 1999). Los significados básicos de 'desposesión' y 'desplazamiento' asociados a estos predicados se funden en una misma conceptualización del estado de cosas cuando el papel semántico de poseedor es a la vez el origen desde el cual se produce un desplazamiento. Ambos verbos organizan esquemas transitivos, con posibles ampliaciones que incluyen junto al objeto un complemento indirecto o/y un complemento adverbial locativo. La proximidad semántica y el paralelismo construccional favorecen el empleo intercambiado de los verbos, pero se comprueba que es en su uso como verbos de apoyo en el que se produce una manifestación más clara de la variación dialectal (Mas 1999, Mas Álvarez 2018). En la lista de sustantivos de la figura 6, *sacar* queda reflejado como verbo con más apariciones en su función de verbo de apoyo y con coocurrencias exclusivas, frente a *quitar*, con los objetos *ventaja, partido, foto, conclusión, provecho, pecho, nota, tajada, basura, punta*, entre otros.

Con la visualización de la diferencia en los perfiles de *quitar* y *sacar* desde los datos del corpus ESLORA[7] se aprecia la confluencia de ambos verbos en la combinación con los colocativos *carné-carnet* y *nota*, una confluencia que apunta sin lugar a dudas al empleo intercambiado —en este caso de *quitar* en lugar de *sacar*— por parte de alguna(s) de las personas participantes en las conversaciones: conseguir el carné de conducir u obtener una nota en un examen son las equivalencias correspondientes a «quitar el carné» (¿usted hace mucho que quitó el carné?) o «quitar una nota» (*estudiaba y quitaba buenas notas*). Igualmente, el ejemplo de «quitar de» con *colegio* representa un uso marcado de *quitar* (*mi padre… la quitó del colegio* [a mi hermana]) frente al no marcado con *sacar*. Con todo, no vemos en esta comparación de perfiles un comportamiento homogéneo en las personas gallegas que participaron en las grabaciones, pues el resto de las coocurrencias presentan empleos no marcados (Mas Álvarez 2018).

[7] El corpus ESLORA (Corpus para el estudio del español oral) contiene 60 horas de entrevistas semidirigidas y 20 horas de conversaciones de hablantes de Galicia grabadas entre los años 2007 y 2015 (ESLORA Versión 1.1.). Para ejemplificar las funcionalidades de *Sketch Engine* se ha volcado el corpus en su totalidad en la plataforma (720193 palabras en 53 documentos).

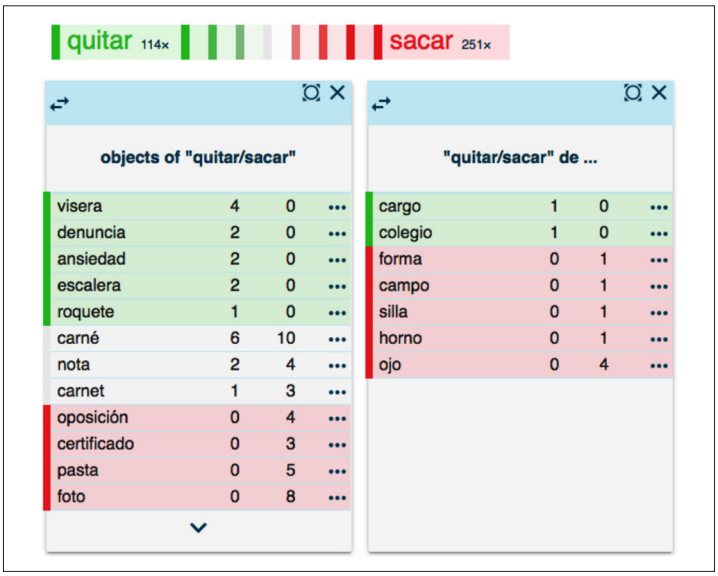


FIGURA 7. *Word Sketch Difference* resumido de los lemas *quitar y sacar* (*ESLORA* en *Sketch Engine*)

Por último, la funcionalidad del catálogo de voces relacionadas, el *Thesaurus*, refleja las palabras que aparecen en contextos similares a los del elemento buscado, es decir, obtenemos un catálogo de voces con un comportamiento similar respecto a las propiedades gramaticales y respecto a las coocurrencias léxicas. La visualización refleja, a través de la cercanía al núcleo, del tamaño de los círculos de color y del dinamismo en la etiqueta, la mayor o menor proximidad de cada elemento del tesauro con la palabra clave, así como la representatividad cuantitativa en el corpus. En la figura 8 se muestra, a modo de ejemplo, el tesauro de «orgulloso».

3.2. El concepto de perfil combinatorio se muestra especialmente adecuado para valorar la asociación preferida entre unidades lingüísticas de cualquier tipo. En semántica cognitiva se emplea para estudiar la proximidad de sinónimos y antónimos, poniendo el énfasis en la asociación léxica. Pero la importancia de los factores gramaticales en la distribución contextual está fuera de cuestión, entre otros motivos porque, al menos desde los trabajos de Sinclair sobre los conceptos relacionados con la coocurrencia de unidades, es patente que el léxico y la gramática de una lengua están mutuamente intrincados (Sinclair 1991).

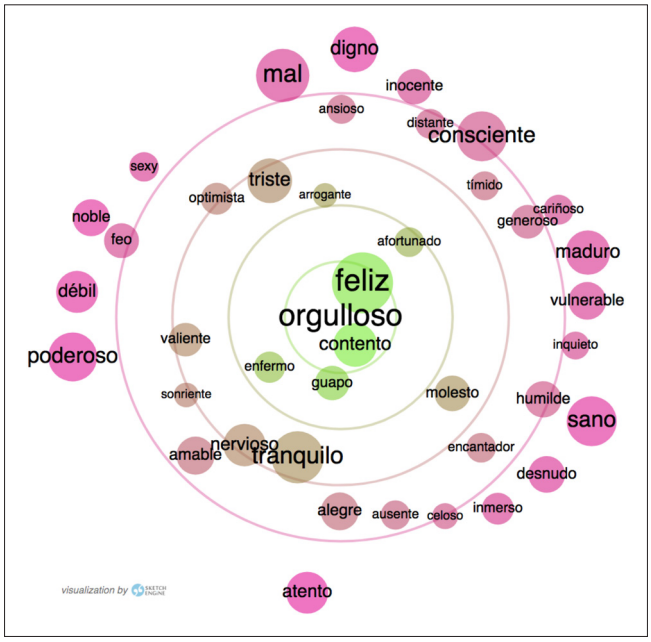


FIGURA 8. Tesauro del lema *orgulloso* limitado a 40 elementos (*Spanish Web 2018 sample* en *Sketch Engine*)

El acercamiento actual más empleado para tratar la atracción mutua entre elementos léxicos y esquemas gramaticales es el propuesto por Stefanowitsch & Gries (2003, 2009), al que los autores llaman *collostruction analysis*. Su método comprende tres pasos (en esta breve caracterización parafraseo a Lehecka 2015: 13-14): primero, el análisis de colexemas mide el grado de atracción o rechazo de un ítem léxico en un hueco del interior de una construcción determinada; segundo, el análisis distintivo de colexemas mide la preferencia de un lema por una construcción determinada sobre otras construcciones similares funcionalmente, y, tercero, el análisis de colexemas covariantes mide el grado de atracción de lemas que aparecen en un hueco de la construcción hacia lemas de otro hueco de la misma construcción.

La atención a hechos léxicos y gramaticales, así como al engranaje entre unidades más abstractas y otras más concretas y su distribución en un corpus son los intereses de un modelo que recurre a técnicas estadísticas refinadas para medir la atracción entre unidades. En Gries (2013) se retomaba el concepto de colocación reivindicando una revisión de las medidas estadísticas de asociación empleadas hasta el momento que permitiera mejorar los resultados en trabajos futuros. A lo largo de los últimos diez años se han publicado

estudios que explotan el método quizá más detallado, según Lehecka (2015: 13), para analizar las interrelaciones entre las propiedades distribucionales y las preferencias de las unidades léxicas: *behavioral profiles analysis* (Gries & Divjak 2009, Gries 2010, Jansegers & Gries 2017).

Según esta metodología, la obtención del perfil combinatorio de una determinada unidad lingüística y su posterior estudio e interpretación requieren seguir cuatro pasos (Gries 2010: 326-328; Jansegers 2017: 134-136):

— Paso 1: la obtención, a partir de un corpus, en forma de concordancias, de todos los ejemplos de los lemas de la(s) palabra(s) que se van a estudiar (o de una muestra aleatoria representativa de ellos).

— Paso 2: el análisis manual y su correspondiente anotación detallada en forma de etiquetas y variables de tantos rasgos como se considere necesario, sean mofológicos, sintácticos, semánticos o de otro tipo. La cantidad de las etiquetas y sus variables dependen de los objetivos del estudio. Por ejemplo, en el trabajo de Jansegers sobre el verbo *sentir* se analizaron 32 rasgos en un total de 153 variables para un corpus que arrojó 1810 ejempos (Jansegers 2017: 336-337).

— Paso 3: la conversión de los datos anotados en el Paso 2 en una tabla de coocurrencias que proporciona las frecuencias relativas de coocurrencia de cada lema o cada significado con cada variable de los rasgos contemplados. Cada columna de la tabla muestra un índice de coocurrencia en porcentaje, el cual constituye, en la concepción de Gries, el perfil combinatorio.

— Paso 4: para terminar, se lleva a cabo la evaluación de los datos de coocurrencia por medio de técnicas estadísticas, como diferencia de porcentajes por pares, acercamientos correlativos o análisis jerárquico de conglomerados. El recurso a esta última técnica ofrece un dendograma o diagrama arbóreo jerárquico en el que aparecen agrupados los elementos más próximos o similares en lo que a su comportamiento distribucional se refiere.

Este método ha sido implementado para estudiar diferentes relaciones léxicas en diversas lenguas (inglés, ruso, francés, español), como la polisemia de verbos concretos, la sinonimia de verbos dentro de una clase semántica, el estudio contrastivo de verbos fásicos (en inglés y ruso) o los adjetivos de tamaño (según la batería recogida en Gries 2010). Más recientemente se ha aplicado a analizar la evolución diacrónica del verbo *sentir* (Jansegers & Gries 2017), incorporando algunas novedades metodológicas, como la adición de una perspectiva dinámica, derivada de la técnica de mapas multidimiensionales a escala (MDS, *multidimensional scaling maps*). Como resultado del

estudio se ofrece una explicación eficiente del proceso seguido por el verbo *sentir* en su especialización semántica y sintáctica en español y, sobre todo, del nacimiento y evolución del marcador discursivo *lo siento*.

4. Consideraciones finales

En el recorrido trazado en este capítulo se han apuntado algunos de los retos a los que se enfrentan los estudios lingüísticos de corte empírico. Obtener el mayor provecho de aquello que solo nos pueden proporcionar los corpus, la frecuencia de aparición de los elementos lingüísticos y las coocurrencias de unidades en forma de concordancias, ha sido el empeño de los acercamientos a los que nos hemos referido en las páginas precedentes. El capítulo va encabezado por tres conocidas citas (Wittgenstein en Kaal & McKinnon 1975: 302, Firth 1957: 11 y Kilgarriff 1997: 108), a las que se podrían sumar algunas más, que expresan una certeza teórica: la preeminencia de los datos obtenidos de la lengua en uso para comprender y explicar los mecanismos lingüísticos y, entre ellos, el más básico, la correspondencia entre construcción-forma y sentido-significado, ese dispositivo apasionante que constituye un entramado multidimensional entre lo que habitualmente llamamos gramática y léxico.

Entender el perfil combinatorio como una esquematización de ese dispositivo nos lleva a algunas disquisiciones de alcance teórico y metodológico, principalmente la referida al trasvase constante entre la conceptualización de las situaciones, la atracción entre elementos léxicos y elementos gramaticales en las construcciones-colocaciones y el significado de las secuencias, puesto que las tres instancias parecen estar mutuamente condicionadas. Como hemos visto, antes de intentar asignar significados, definiciones o incluso traducciones a un elemento léxico, primero se deberán identificar los varios patrones sintácticos y colocacionales diferentes en los que ese elemento se presenta habitualmente; pero, en el curso de tal identificación, asoman las propiedades contextuales más relevantes, que serán las que permitan confirmar cuáles son los parámetros destacados que es necesario tener en cuenta.

Para establecer la nómina de parámetros destacados se hace imprescindible dirigir la atención fuera del límite oracional siempre que se muestren como significativas las propiedades discursivas. Esta ampliación del contexto es inexcusable en el estudio de la lengua oral. Por otra parte, ya se ha insistido sobre ello, moverse en diversos niveles de generalización significa no desatender el detalle, que puede constituir la clave de diferenciación en acercamientos contrastivos: conocer las unidades léxicas concretas que entran en el

perfil combinatorio, su frecuencia y su grado de atracción o rechazo respecto a una construcción, otro lexema o una propiedad gramatical o discursiva se revela como esencial para el establecimiento de los fenómenos idiomáticos.

Desde el punto de vista metodológico, lidiar con el tamaño de los corpus, su diseño, el acceso a las concordancias —no siempre tan rápido y sencillo como sería deseable— y los problemas asociados a la obtención de una muestra aleatoria de los ejemplos y su traslado a un entorno de trabajo resulta aún bastante penoso. Quizá no estemos ya tan lejos de conseguir una automatización fiable de los procesos para identificar las categorías, al menos en aquellas etiquetas menos necesitadas de desambiguación manual. El ritmo acelerado al que nos tienen acostumbradas las mejoras tecnológicas con frecuencia parece ir en contra de lo que semejarían avances naturales. Las herramientas que ofrece *Sketch Engine* son óptimas, pero resultan aún muy mejorables, pues los módulos de lematización y etiquetado son demasiado básicos, al menos para el español (para otras lenguas ni siquiera existen). Por todo ello, no parece fácil escapar del análisis manual minucioso ejemplo a ejemplo, a pesar de que en aras de la representatividad estemos abocados a consultar grandes colecciones de concordancias.

En cuanto a las técnicas estadísticas implicadas en la obtención de los perfiles combinatorios, son muchas y diversas, y aún no parecen lo suficientemente consolidadas. Los efectos que produce la ley de Zipf en las pruebas, dependientes del tamaño de los corpus, exigen evaluar los resultados que arrojan. Es necesario conjugar la aplicación de métodos cuantitativos y cualitativos: los primeros son necesarios, pero no suficientes; los segundos se benefician de la generalización y la visualización sinóptica que brindan los índices de frecuencia y asociación de las coocurrencias.

Corpus

ADESSE. *Base de datos de verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español*. Universidade de Vigo. adesse.uvigo.es

BDS. *Base de Datos Sintácticos del español actual*. Universidade de Santiago de Compostela. Versión 3.5.3. www.bds.usc.es

ESLORA. *Corpus para el estudio del español oral*. Versión 1.1. de marzo de 2018. eslora.usc.es

SKETCH ENGINE. *Language corpus management and query system*. www.sketchengine.eu

Referencias bibliográficas

Alonso-Ramos, Margarita (2009): «Hacia un nuevo recurso léxico: ¿fusión entre corpus y diccionario?», in Pascual Cantos Gómez & Aquilino Sánchez Pérez (eds.): *A survey of corpus-based research*, Murcia: Asociación Española de Lingüística de Corpus, pp. 1191-1207.

Baider, Fabienne & Georgeta Cislaru (eds.) (2013): *Cartographie des émotions: propositions linguistiques et sociolinguistiques*. Paris: Presses Sorbonne Nouvelle.

Blumenthal, Peter (2002): «Profil combinatoire des noms: synonymie distinctive et analyse contrastive», *Zeitschrift für französische Sprache und Literatur* 112/2, pp. 115-138. www.jstor.org/stable/40618538

Firth, John Rupert (ed.) (1957): *Studies in linguistic analysis*. Oxford: Blackwell.

García-Miguel, José M. (2014): «El perfil combinatorio de los verbos en ADESSE: polisemia y parasinonimia de verbos de competición», in Yuko Morimoto (ed.): *Léxico, didáctica y nuevas tecnologías,* Anexos, *Revista de Lexicografía* 29, pp. 11-37.

García-Miguel, José M. & Francisco J. Albertuz (2005): «Verbs, semantic classes and semantic roles in the ADESSE project», in Erk Katrin, Alissa Melinger & Sabine Schulte im Walde (eds.): *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbrücken, 28 febrero - 1 marzo 2005. adesse.uvigo.es/textos/saarb05.pdf

Gries, Stefan Th. (2010): «Behavioral profiles: a fine-grained and quantitative approach in corpus-based lexical semantics», *The Mental Lexicon* 5/3, pp. 323-346. https://doi.org/10.1075/ml.5.3.04gri

Gries, Stefan Th. (2013): «50 something years of work on collocations. What is or should be next?», *International Journal of Corpus Linguistics* 18/1, pp. 137-165. https://doi.org/10.1075/ijcl.18.1.09gri

Gries, Stefan Th. & Dagmar Divjak (2009): «Behavioral profiles: a corpus-based approach towards cognitive semantic analysis», in Vyvyan Evans & Stéphanie Pourcel (eds.): *New directions in cognitive linguistics*. Amsterdam: John Benjamins, pp. 57-75. https://doi.org/10.1075/hcp. 24.07gri

Hanks, Patrick (1996): «Contextual dependency and lexical sets», *International Journal of Corpus Linguistics* 1/1, pp. 75-98. https://doi.org/10.1075/ijcl.1.1.06han

Hanks, Patrick (2013): *Lexical analysis. norms and exploitations*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262018579.001.0001

Hanks, Patrick (en línea): *A Pattern Dictionary of English Verbs*. [PDEV] pdev.org.uk

Jansegers, Marlies (2017): *Hacia un enfoque múltiple de la polisemia: un estudio empírico del verbo multimodal «sentir» desde una perspectiva sincrónica y diacrónica*. Berlin: de Gruyter. https://doi.org/10.1515/9783110476972

Jansegers, Marlies & Stefan Th. Gries (2017): «Towards a dynamic behavioral profile: a diachronic study of polysemous *sentir* in Spanish», *Corpus Linguistics and Linguistic Theory* 13/1, pp. 1-43. https://doi.org/10.1515/cllt-2016-0080

Kaal, Hans & Alastair McKinnon (eds.) (1975): *Concordance to Wittgenstein's* Philosophische Untersuchungen *compiled*. Leiden: E. J. Brill.

Kilgarriff, Adam (1997): «'I don't believe in word senses'», *Computers and the Humanities* 31/2, pp. 91-113. https://doi.org/10.1023/A:1000583911091

Kilgarriff, Adam (2005): «Putting the corpus into the dictionary», *Proceedings of the MEANING Workshop*. Trento. February 2005. https://bit.ly/2PNijYW

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel (2014): «The Sketch Engine: Ten years on», *Lexicography ASIALEX* 1/7, pp. 7-36. https://doi.org/10.1007/s40607-014-0009-9

Lehecka, Tomas (2015): «Collocation and colligation», *Handbook of pragmatics*. Amsterdam: John Benjamins. https://doi.org/10.1075/hop.19.col2

Mas, Inmaculada (1999): «El intercambio de los verbos *sacar* y *quitar* en el castellano de Galicia», in Rosario Álvarez Blanco & Dolores Vilavedra (eds.): *Cinguidos por unha arela común: homenaxe ó profesor Xesús Alonso Montero*, Vol. I, Santiago de Compostela: Universidade de Santiago de Compostela, Servizo de Publicacións e Intercambio Científico, pp. 655-675.

Mas, Inmaculada & Guillermo Rojo (2004): «Design, Construction and Exploitation of the Contemporary Spanish Syntactic Database», in Ulrich Engel & Meike Meliss (eds.): *Dependenz, Valenz und Wortstellung*, München: Iudicium, pp. 101-108.

Mas Álvarez, Inmaculada (2018): «Los verbos *quitar* y *sacar*: sus perfiles combinatorios a partir de datos de corpus», comunicación presentada al *XIII Congreso Internacional de Lingüística Xeral*, Universidade de Vigo, Vigo, 13-15 de xuño de 2018.

Mellado, Carmen & Belén López Meirama (2017): «El fraseologismo 'por momentos': principales valores semánticos y algunos apuntes diatópicos», *RILCE: Revista de filología hispánica* 33/2, pp. 648-670. https://doi.org/10.15581/008.33.2.648-70

Rojo, Guillermo (2001): «La explotación de la Base de Datos Sintácticos del español actual (BDS)», in Josse De Kock (ed.): *Gramática española: enseñanza e investigación*, Salamanca: Universidad de Salamanca, pp. 255-286. www.bds.usc.es/index.html?url=masinfo.html

Rojo, Guillermo (2015): «Sobre los antecedentes de la lingüística do corpus», in Alfredo Álvarez Menéndez (ed.): *Studium grammaticae: homenaje al profesor José Antonio Martínez*, Oviedo: Universidad de Oviedo, pp. 675-689.

Rojo, Guillermo & Victoria Vázquez Rozas (2014): «Sobre las formas en *-ra* en el español de Galicia», in Andrés Enrique-Arias, Manuel J. Gutiérrez, Alazne Landa & Francisco Ocampo (eds.): *Perspectives in the study of spanish language variation: papers in honor of Carmen Silva-Corvalán*. Santiago de Compostela: Universidade de Santiago de Compostela, pp. 237-270. https://doi.org/10.15304/va.2014.701

Sinclair, John (1991): *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stefanowitsch, Anatol & Stefan Th. Gries (2003): «Collostructions: investigating the interaction of words and constructions», *International Journal of Corpus Linguistics* 8/2, pp. 209-243. https://doi.org/10.1075/ijcl.8.2.03ste

Stefanowitsch, Anatol & Stefan Th. Gries (2009): «Corpora and grammar», in Anke Lüdeling & Merja Kytö (eds.): *Corpus linguistics: an international handbook*, vol. 2. Berlin: Mouton de Gruyter, pp. 209-243. https://doi.org/10.1515/9783110213881.2.933

FUNCIONES PRAGMÁTICAS EN EL PORTUGUÉS BRASILEÑO: UN ENFOQUE DISCURSIVO-FUNCIONAL

Pragmatic functions in Brazilian Portuguese: a Functional Discourse Grammar account

Hella Olbertz
Universidad de Amsterdam

Resumen

El presente trabajo versa sobre las funciones pragmáticas de tópico y foco en los referentes del sujeto gramatical de cláusulas transitivas en el portugués hablado en Brasil, en comparación con las variedades europeas del español y del portugués. El portugués brasileño está desarrollando un marcador de tópico innovador, que consiste en un pronombre personal que sigue inmediatamente al sintagma nominal sujeto, pero esta lengua tiene posibilidades muy limitadas para expresar la función de foco. El portugués europeo carece de marcas específicas para ambas funciones pragmáticas. El español, por otra parte, difiere de las dos variedades del portugués por disponer de una expresión de foco consistente en la posición posverbal del referente del sujeto, pero no tiene expresión sistemática de tópico. En este estudio se describe el origen del marcador de tópico en el portugués brasileño, se explica su funcionamiento, y se ofrece una comparación sistemática de las posibilidades de expresión de las dos funciones pragmáticas en las variedades europeas de estas lenguas iberorrománicas y en el portugués brasileño. El análisis se circunscribe a las cláusulas verbales transitivas cuyos dos argumentos tienen referentes de tercera persona.

Palabras clave: tópico, foco, sujeto, cláusula transitiva, pronombre personal

Abstract

This study concerns the possibilities of expressing the pragmatic functions of topic and focus on subject referents of transitive clauses in Portuguese as spoken in Brazil,

compared with the European varieties of Spanish and Portuguese. Spoken Brazilian Portuguese is developing an innovative topic marker for referents with subject function, which consists of the redundant use of a personal pronoun that immediately follows the subject NP, but it has only very restricted possibilities of expressing the focus function. European Portuguese lacks specific expressions for both pragmatic functions. Spanish differs from Portuguese in being able to express focus by placing the grammatical subject in clause final position; but it does not have an analogous expression of the topic function. The present paper describes the origin of the topic marker in Brazilian Portuguese, explains its functionality, and ends with a comparison of the possible expressions of topic and focus in the European varieties of the two Ibero-Romance languages and Brazilian Portuguese. The paper is concerned exclusively with transitive verbal expressions in which both arguments have third person referents.

Keywords: topic, focus, subject, transitive clause, personal pronoun

1. Introducción

Aunque el español es una lengua con orden básico de constituyentes sujeto-verbo-objeto (SVO), el orden de los argumentos del verbo puede variar, así que también es posible que el objeto preceda al verbo y el sujeto siga al verbo (OVS)[1]. El desvío del orden básico puede tener diferentes motivos[2], uno de los cuales es de naturaleza pragmática: el constituyente con más valor informativo tiende a situarse al final del enunciado (Zubizarreta 1999: 4232-4234; Martínez Caro 2007: 129-130; Hannay & Martínez Caro 2008: 38-41; Leonetti 2011: 338):

(1) Los de la revista de Petrikorena dicen tener pruebas de que a su director lo mató la Guardia Civil. (CREA, Alfonso Rojo, *Matar para vivir*. Barcelona: Plaza y Janés, 2002)

En este ejemplo, el sintagma nominal (SN) que se refiere a la presunta responsable de la muerte del director de la revista, la Guardia Civil, ocupa la posición posverbal y final de la cláusula, lo cual se debe a su función de foco, más concretamente de «foco informativo»[3].

[1] Agradezco a María José Rodríguez Espiñeira, Lachlan Mackenzie, Victoria Vázquez, Leo Lemmers y un revisor anónimo sus comentarios valiosos a versiones previas de este trabajo. Los errores que queden son míos.

[2] El aspecto más conocido que determina el orden de los constituyentes es el «peso» de los argumentos o elementos adverbiales: los constituyentes complejos tienden a seguir a los constituyentes menos complejos según el *Gesetz der wachsenden Glieder* de Behagel (1932) (citado por Dik 1997: I, 404).

[3] Como gran parte de la bibliografía reciente sobre la estructura informativa es de corte generativo, cabe aclarar que se trata aquí del «foco informativo» y no del así llamado «foco de polaridad» o *verum*

En portugués un procedimiento análogo no es posible con este ejemplo, como se puede apreciar en la siguiente traducción literal de (1), que resulta bastante extraña.

(1) a. ≠ *dizem ter provas de que seu diretor matou a Guarda Civil*

Según la única interpretación posible, el director de la revista es el referente agentivo que mató a la Guardia Civil[4]. Este ejemplo demuestra que el portugués no puede marcar el foco por medio del orden de constituyentes en las predicaciones transitivas cuando los dos argumentos tienen referentes humanos, porque carece de una marca diferencial de objeto análoga a la preposición *a* de objeto del español. Cuando el argumento paciente de una predicación transitiva no tiene referente humano, el portugués sí admite la posposición del argumento agente, pero solamente si los dos argumentos ocupan la posición posverbal, como en el siguiente ejemplo inventado del portugués europeo:

(2) – *Quem comeu a sopa?*
– *Comeu a sopa a Maria*. (Martins 2011: 135)
'– ¿Quién comió la sopa?
– Comió la sopa María.'

En (2) la posposición del primer argumento es posible por ser el único referente humano que, como tal, puede funcionar como agente; por lo tanto, la marca diferencial de objeto es irrelevante en este contexto.

También existen expresiones de tópico en las dos lenguas, limitadas, sin embargo, a los referentes de argumentos no-sujetos:

(3) 2. ya:// para la gente mayor ¿tú crees que está cubierto o cómo lo ves// (e:)// este: ...?
1. a la gente mayor la cuidan// a la gente mayor la cuidan les interesan// son muchos votos
(AdH 27 media 49 masc)[5]

focus del tipo *Algo debe saber* que suele aparecer al comienzo del enunciado (Leonetti 2011: 340-341; para más detalles *cfr*. Bosque & Gutiérrez-Rexach 2009, capítulo 11).

[4] Para el portugués europeo, Lachlan Mackenzie sugiere usar *é que* para la expresión de foco. Sin embargo, como la variabilidad del uso del portugués *é que* y del español *es que* es tan amplia, considero preferible no discutirlo en el marco del presente capítulo (para el español, véase por ejemplo Fuentes Rodríguez 1997).

[5] AdH = PRESEEA de Alcalá de Henares; Iboruna = *Banco de dados Iboruna: amostras eletrônicas do português falado no interior paulista*. Las indicaciones de las fuentes de ejemplos consisten en el

(4) *Doc.: foi nessa época que a senhora conheceu o seu marido?*
Inf.: nã:o meu marido eu conheci: depois de muito/ muita ida:de
(Iboruna 144 media 58 fem)[6]
'Documentalista: ¿fue en esa época cuando usted conoció a su marido?
Informante: no, a mi marido lo conocí cuando ya era más mayor'

Como el portugués no dispone ni de una marca diferencial de objeto ni de duplicación de clítico (*la* preverbal en el ejemplo 3), marcas sintácticas que conjuntamente sirven para indicar la función sintáctica del SN antepuesto en español, la marcación del tópico por medio de la variación del orden de palabras es infrecuente en portugués[7]. Nótese, además, que en el ejemplo (4) el referente del sujeto es de primera persona, por lo cual no puede ofrecer ambigüedad. Es improbable que en portugués se encuentren ejemplos análogos a (3), donde los dos argumentos hacen referencia a la tercera persona.

Por otra parte, en el portugués hablado en Brasil se está desarrollando una marca de tópico para el referente del sujeto, que consiste en el uso del pronombre personal que sigue inmediatamente, es decir sin pausa intermedia, al sintagma nominal en función de sujeto:

(5) *a cabeleireira ela tem que ter uma tesoura que chama: ... ai como chama aquela tesoura gente?...* (Iboruna 72 primaria 28 fem)
'la peluquera tiene que tener unas tijeras que se llaman ... ay ¿cómo se llaman aquellas tijeras eh?...

Al ser sintácticamente redundante, el pronombre de sujeto se reinterpreta como marca de tópico (*cfr*. p. ej. Duarte 1995: 100-124; Barbosa, Duarte & Kato 2005).

---

número o el título de la entrevista, al que siguen el nivel de instrucción (primaria, media, superior), la edad y el sexo (fem, masc) del informante.

[6] Las transcripciones originales reflejan fielmente las características fonológicas del habla del interior del Estado de São Paulo. Se han eliminado muchas de ellas en los ejemplos por ser irrelevantes para este estudio. También están omitidas las marcas de habla simultánea y los sonidos paralingüísticos emitidos por el documentalista para animar al informante a que siga hablando. Se han mantenido las marcas de alargamiento (:), las pausas (...), las mayúsculas para marcar el énfasis y algunas simplificaciones del tipo *num* por *não* en posición preverbal, *tá* por *está, c'o* por *com o* y similares.

[7] Barbosa, Duarte & Kato (2005: 21) sostienen que en el portugués la posición posverbal del sujeto se limita a cláusulas inacusativas, como la del siguiente ejemplo, en el cual el verbo *ficar* es inacusativo y los referentes de sujeto pospuestos son *o Antunes...* y *o*: *Edinho.*

(i) *ficou agora pro segundo turno [...] aqui pra prefeito o Antunes... e o*: *Edinho* (Iboruna 152 superior 76 fem)
'quedó ahora para la segunda ronda [...] aquí para alcalde el Antunes ... y el Edinho'

Los objetivos del presente trabajo son (i) describir desde una perspectiva funcional el desarrollo de este fenómeno, (ii) explicar su comportamiento dentro de un corpus oral, para (iii) presentar una comparación sistemática sobre la expresión de las funciones de foco y de tópico en las tres variedades. Sin embargo, antes de entrar en estos detalles, hace falta explicar lo que entiendo por tópico, foco y otras funciones pragmáticas, y mostrar cómo se relacionan estas con las propiedades semánticas y sintácticas de los predicados y sus argumentos.

Este capítulo está organizado así: en la sección 2 presentaré los fundamentos teóricos de mi análisis; la sección 3 estará dedicada al origen de la marca de tópico del portugués brasileño y la sección 4 a su funcionamiento en el habla. En la sección 5 usaré los resultados de los apartados 3 y 4 para comparar el comportamiento del portugués de Brasil con el del español, por un lado, y el del portugués europeo, por otro. La sección 6 contiene las conclusiones.

Este trabajo consiste en un análisis esencialmente cualitativo de corpus lingüísticos orales. Los corpus españoles (AdH, PRESEEA de Alcalá de Henares, 443 533 palabras) y brasileños (Iboruna, interior del Estado de São Paulo, 767 943 palabras) son comparables por ser entrevistas sociolingüísticas con informantes clasificados sistemáticamente según el nivel de instrucción, la edad y el sexo. Además, los dos se pueden considerar como representativos de las variedades metropolitanas del español y portugués brasileño[8], respectivamente. Una diferencia esencial entre los dos reside en que el corpus brasileño incluye hablantes a partir de los 7 años de edad. El corpus del portugués europeo abarca la parte informal portuguesa de C-Oral-Rom (COR-P, 137 243 palabras), que integra tanto entrevistas como conversaciones más informales de más de dos participantes. Como el COR-P es relativamente pequeño, se añadió una parte europea del corpus *Português Falado* (PF-P, 26 160 palabras), entrevistas grabadas en los años 90 del siglo XX. Nótese que para los análisis cuantitativos se hace uso de datos cuantitativa y cualitativamente comparables.

2. Fundamentos teóricos: gramática discursivo-funcional

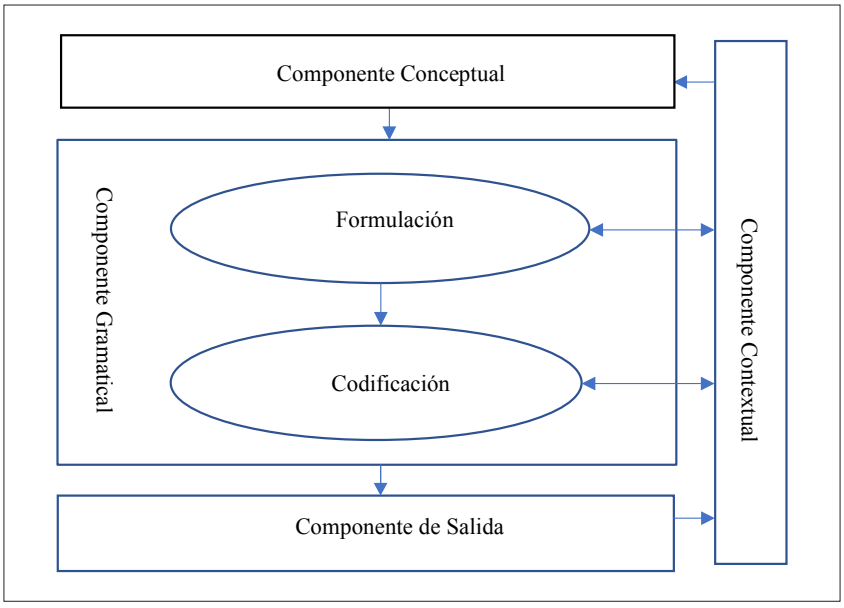
En la introducción a este trabajo he usado los términos de tópico y foco, de agente y paciente, y de sujeto y objeto sin indicar a qué conceptos remiten. Para comprender su naturaleza y mostrar sus diferencias, describiré breve-

[8] Queda excluida la variedad del extremo sur del Brasil, que conserva la segunda persona del singular con el pronombre *tu* como forma de tratamiento informal.

mente las ideas fundamentales de la gramática discursivo-funcional en la medida en que son relevantes para este trabajo. Para una descripción detallada de la teoría véanse Hengeveld & Mackenzie (2008; 2011).

La gramática discursivo-funcional (GDF) es un modelo formalizado que pretende dar cuenta de la estructura de la lengua tal como se usa; por tanto, la unidad básica de la GDF no es la cláusula sino el enunciado. La estructura del modelo está basada en los estudios tipológicos, lo cual implica que las lenguas individuales «se toman en serio», es decir que no se presuponen universales lingüísticos *a priori*, sino que el sistema debe servir para cada lengua individual y, si resulta no ser adecuado para tal descripción, habrá de ser adaptado.

La GDF forma parte de una teoría general de la comunicación verbal y, como tal, interactúa con tres componentes extralingüísticos. El cuadro 1 (adaptado de Hengeveld & Mackenzie 2008: 6) proporciona una visión general de cómo funciona.



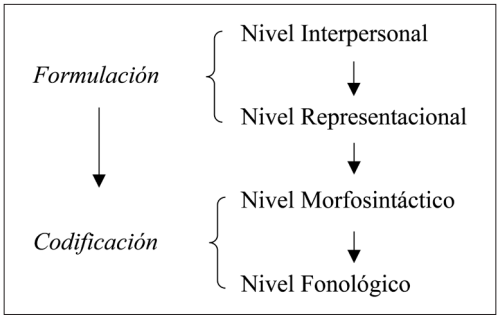
CUADRO 1. Gramática discursivo-funcional y la interacción verbal

El rectángulo central es una representación simplificada de la estructura de la gramática, el único componente lingüístico propiamente dicho del total, puesto que los demás componentes son de naturaleza extralingüística. El único componente con el que la gramática interactúa bidireccionalmente

es el componente contextual, representado por el rectángulo de la derecha, que almacena el contexto lingüístico y extralingüístico de la comunicación y el conocimiento del mundo del hablante. La interacción bidireccional se debe al hecho de que todos los aspectos lingüísticos de la comunicación pueden llegar a ser objeto de enunciados metalingüísticos. El componente conceptual, representado por el rectángulo superior, es responsable de la formación de la intención comunicativa que, a su vez, está influenciada por los datos almacenados en el componente contextual (García Velasco 2014). El componente conceptual es «la fuerza motriz tras el componente gramatical en su totalidad» (Hengeveld & Mackenzie 2011: 7). El componente de salida, finalmente, transforma la salida del componente gramatical en habla, escritura o signos (en las lenguas de signos) producidos por el individuo con todas sus propiedades idiosincrásicas. El contenido del componente de salida, a su vez, entra en el componente contextual ya que forma parte de la información contextual potencialmente relevante para enunciados futuros (García Velasco 2014).

Es importante notar que el componente gramatical solamente da cuenta de las realizaciones estrictamente lingüísticas de las intenciones del hablante. El componente gramatical consiste en dos componentes básicos en un orden jerárquico, la formulación y la codificación, representados como elipses en el cuadro 1. La formulación es responsable de los contenidos interpersonales y semánticos del enunciado, mientras que la codificación es de naturaleza morfosintáctica y fonológica y no afecta al contenido del enunciado. Ambos representan la realización dinámica del enunciado; el orden mutuo de formulación y codificación es lógico ya que, desde una perspectiva funcional, la lengua se considera en primer lugar como medio de comunicación y no como un «conjunto infinito de oraciones».

En el cuadro 2 se representa de manera simplificada la estructura interna de formulación y codificación.



Cuadro 2. Formulación y codificación del enunciado en GDF

En el Nivel Interpersonal se da cuenta de los aspectos lingüísticos estratégicos de la comunicación, mientras que el Nivel Representacional recoge todos los aspectos descriptivos de la comunicación. Dicho de otra manera, el Nivel Interpersonal es responsable de lo que *hace* el hablante y el Nivel Representacional es responsable de lo que *dice*.

La entidad básica del Nivel Interpersonal es el «acto discursivo» (A_1), cuya estructura básica es la siguiente:

(6) (A_1: [(Ilocución) (P_1)$_{Hablante}$ (P_2)$_{Oyente}$ (C_1: [(Adscripción$_1$)$_\varphi$ (Referente$_1$)$_\varphi$] (C_1))] (A_1))[9]

Es importante tener en cuenta que los actos discursivos no necesitan equivaler a una cláusula, sino que se definen como unidad mínima de comunicación (Kroon 1995: 65-66). Pueden consistir en un saludo, una felicitación o cualquier otra forma de enunciado. Los interlocutores (P) están presentes dentro del acto discursivo para dar cuenta, entre otros, de la formalidad de la relación entre ellos (por medio de un operador honorífico) y de la autorreferencia al hablante en actos discursivos performativos. La entidad mayor de la que forman parte los actos discursivos se llama «movimiento» (M), que en la interacción oral suele equivaler a un turno y en la lengua escrita a un párrafo.

Lo que nos interesa en la estructura (6) es el contenido comunicado (C_1), que contiene los «subactos» de adscripción (T_1) y referencia (R_1), ya que a estos se les atribuyen las funciones pragmáticas (φ). La tipología de las funciones pragmáticas no opone el tópico al foco, sino que ambos son los elementos marcados de sendas oposiciones bimembres:

(7) Foco – Trasfondo
Tópico – Comentario
Contraste – Solapamiento

Generalmente solo los elementos marcados de estas oposiciones, foco, tópico y contraste, suelen tener una expresión lingüística. Como no existe oposición entre ellas, estas tres funciones pragmáticas se pueden combinar. En (8) se ejemplifica tanto la combinación de tópico y foco como la combinación de tópico y contraste:

[9] El formalismo usado en la GDF es análogo al de la lógica de predicados: (α: *f* (α)) se lee como «una variable α tal que *f* es una propiedad de α». Es decir que existe una relación «jerárquica» entre α y *f*, tal que α tiene *f* dentro de su alcance. Los corchetes en las representaciones encierran relaciones entre entidades en una relación no jerárquica. Las abreviaturas usadas se incluyen en una lista al final de este capítulo.

(8) mi hijo compró el piso a unos que se iban a casar/ y estaba:/ sin estrenar/ y: ocho días antes de:- de la boda// regañaron/ y ella se ha casado con otro/ y él está en Canarias/
(AdH 52 primaria 62 fem)

En este ejemplo las funciones pragmáticas se marcan mediante el uso del pronombre personal: *ella* representa un «tópico nuevo», que en la GDF se representa como la combinación de tópico y foco (TopFoc), y el pronombre *él* se contrasta con *ella* y por tanto se analiza como «tópico contrastivo», es decir una combinación de tópico y contraste (TopContr).

En los estudios sobre las funciones informativas se suelen considerar las expresiones «dislocadas a la izquierda» como tópico (frecuentemente llamado *hanging topic* o bien «tema vinculante») y las «dislocadas a la derecha»[10] como foco. El siguiente ejemplo ilustra el primero de estos fenómenos:

(9) a. Dinero, todo el mundo lo necesita (Zubizarreta 1999: 4220)
b. *O incêndio, os bombeiros controlaram-no* (Andrade Peres & Móia 1995: 32)
'El incendio, los bomberos lo controlaron'

La perspectiva de la GDF es que este tipo de elementos separados por una pausa del resto no forman parte de un acto discursivo único, sino que son actos discursivos dependientes del principal, y que la relación entre los dos se expresa por medio de funciones retóricas. En el caso de (9) la función de los actos discursivos dependientes, *dinero* y *o incêndio* sería la de «orientación». En el ejemplo (10), el sintagma nominal «dislocacado a la derecha», *esses candeeiros*, es un acto discursivo con la función retórica denominada «corrección», aunque quizás sea más apropiada la etiqueta «explicación».

(10) *nem sequer... dão muita luz, esses candeeiros*. (PF-P «Saber vender» primaria 46 fem)
'ni siquiera ... dan mucha luz, esos quinqués'

El Nivel Representacional es el lugar de los contenidos proposicionales, que se evalúan según su valor de verdad, y de los estados de cosas con predicados verbales, adjetivales o nominales y argumentos a los que se atribuyen funciones semánticas del tipo agente (*Actor*), paciente (*Undergoer*) y receptor.

En el Nivel Morfosintáctico se determina el orden de constituyentes según los patrones específicos de cada lengua, sin que se haga uso de movimientos

[10] Como se verá más adelante, la sintaxis de la GDF no hace uso de transformaciones, así que no existen «dislocaciones» en esta teoría. Se emplea este término aquí por ser de uso común.

u otro tipo de transformaciones. El orden de constituyentes puede depender de factores pragmáticos (interpersonales) y semánticos (representacionales), pero también de factores puramente morfosintácticos, entre ellos la función sintáctica de sujeto, que se puede asignar al agente o paciente y en algunas lenguas incluso al receptor (p. ej. en inglés). Lo que es importante para el presente trabajo es la manera de marcar el sujeto. En las lenguas románicas y germánicas hay tres opciones[11]: se puede realizar (i) por medio del pronombre solo, sin que haya concordancia (como las lenguas escandinavas continentales), (ii) por medio de la desinencia verbal cuando el componente contextual proporciona la identidad del referente (como en la mayoría de las lenguas románicas), (iii) por medio de la combinación obligatoria de los dos, pronombre y desinencia (como en francés y alemán). Al primer caso (solo pronombre) lo denominaré «referencia nominal única»; el segundo (solo desinencia – si el contexto lo permite) se ha designado como «concordancia contextual» (Hengeveld 2012) y el tercero (pronombre y desinencia) se suele llamar «concordancia sintáctica» (Siewierska 2004: 126).

En el Nivel Fonológico la unidad que suele corresponder al acto discursivo es la frase entonativa (*Intonational Phrase*, IP). Este nivel es importante para nuestros propósitos porque refleja la diferencia entre la construcción de dos actos discursivos en el ejemplo (9) y la de tópico, ilustrada en (5): en el primer caso hay dos frases entonativas y en el segundo solo una (Hengeveld & Mackenzie 2008: 433).

Llegados a este punto, quisiera subrayar algunos aspectos de esta breve presentación teórica que resultan relevantes para el tema analizado, la expresión morfosintáctica de las funciones pragmáticas. En primer lugar, he mostrado cuáles son las funciones pragmáticas y cómo se usan, y que es preciso distinguir las funciones pragmáticas dentro de un acto discursivo único y las funciones retóricas que relacionan los actos discursivos dependientes con los independientes, una diferencia que se refleja de modo sistemático en la codificación fonológica. En segundo lugar, he llamado la atención sobre el hecho de que estas funciones forman parte del Nivel Interpersonal, cuyo papel es exclusivamente estratégico, no relacionado con los valores de verdad de las proposiciones, que en mi aproximación pertenecen a un nivel distinto de la lengua, el Nivel Representacional. En tercer lugar, he explicado la función limitada que posee la morfosintaxis en este enfoque, pues no tiene repercusión sobre los *contenidos* que se expresan, si bien tiene efectos importantes sobre la *manera* en que se expresan.

[11] En esta clasificación empleo la palabra *pronombre* para referirme al pronombre personal de sujeto, cuyo lugar puede ser ocupado alternativamente por un SN léxico o un nombre propio.

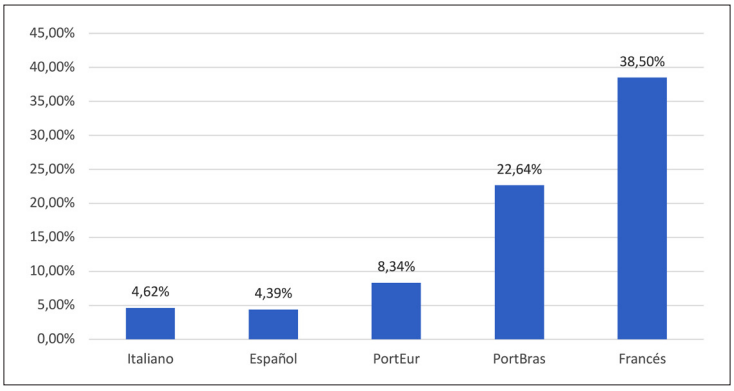
3. El desarrollo del marcador de tópico en el portugués brasileño

Como he indicado en la introducción de este trabajo, el marcador de tópico innovador del portugués de Brasil (PB) consiste en un pronombre personal tónico que sigue inmediatamente al SN sujeto, sin que haya una pausa entre las dos palabras. Este fenómeno no es específico del habla brasileña (11), sino que también se presenta en el francés hablado (12):

(11) *meu filho ele costumava conversar com ela pelo*: *... pela internet*
(Iboruna 78 media 31 fem)
'mi hijo solía conversar con ela por el ... por la internet'

(12) *donc le patron il peut dire à l'occasion: «Euh bon j'ai plus besoin de vous, vous pouvez partir.»* (CSF 16 primaria 55 masc)
'entonces el jefe puede decir en tal momento: «Bueno ya no le necesito a usted, puede irse.»'

Para que tal fenómeno se produzca, es necesario que haya una cierta frecuencia en el uso del pronombre personal, lo cual obviamente sucede en el caso del francés por tener concordancia sintáctica, pero no es esperable en el portugués porque pertenece al tipo de lenguas con concordancia contextual. Sin embargo, los datos del cuadro 3, recopilados a partir de C-Oral-Rom (Cresti & Moneglia 2005) y C-Oral-Brasil (Raso & Mello 2012) muestran que el uso del pronombre de sujeto del PB es atípicamente alto para una lengua de concordancia contextual.



Cuadro 3. Sujetos pronominales en cinco lenguas y variedades románicas[12]

[12] Fueron contados los pronombres de sujeto en relación al número de verbos finitos. Se contaron exclusivamente los casos en los cuales el pronombre precedía inmediatamente al verbo finito, admi-

Mientras que la escasa diferencia existente entre el italiano y el español en el cuadro 3 probablemente se debe a una coincidencia[13], el uso del sujeto pronominal es un poco más frecuente en el portugués europeo (PE), y en el de Brasil se encuentra entre los valores del PE y del francés. Así que parece que el PB está evolucionando en la dirección del francés, es decir hacia una lengua con concordancia sintáctica.

Considérese el siguiente grupo de ejemplos construidos:

(13) a. Español
Su hijo (R_i) vive en Suiza donde él (R_j) trabaja en una fábrica.
b. Francés
Leur fils (R_i) habite en Suisse où il (R_i) travaille dans une usine.
c. Portugués brasileño
O filho deles (R_i) mora na Suiça onde ele ($R_{i/j}$) trabalha numa fábrica.

En español, el uso de pronombre de sujeto se limita a los casos pragmáticamente y/o contextualmente motivados; tal motivación puede consistir en la reintroducción de un referente mencionado anteriormente en contraste con el referente *su hijo* (es decir, con función de tópico contrastivo), lo cual sería la interpretación más lógica de *él* en (13a). En (13b) por otra parte, no cabe duda de que *il* tiene el mismo referente que *leur fils*, por ser obligatoria la especificación del sujeto en francés. El PB ocupa una posición intermedia, el pronombre *ele* en (13c) puede leerse de acuerdo con la concordancia contextual, es decir de la misma manera que *él* en (13a), o alternativamente, de acuerdo con la concordancia sintáctica como pronombre correferencial con *o filho deles*, análogo a la interpretación de *il* en (13b).

En las tres subsecciones que siguen describiré brevemente a qué obedece esta situación ambigua del PB (3.1), después mostraré cómo se ha desarrollado la función de marca de tópico en esta variedad (3.2) y terminaré discutiendo un posible argumento en contra de este análisis (3.3).

---

tiéndose un clítico entre pronombre y verbo. Los números absolutos de sujetos pronominales por verbo finito son los siguientes: en italiano 2 953 de un total de 64 797; en español 2 368 de 53 939; en el portugués europeo 4 825 de 58 167; en el portugués brasileño 8 242 de 36 398; y en francés 18 461 de 47 954. El número de pronombres de sujeto relativamente bajo del francés se debe tanto al número elevado de inserciones adverbiales tales como *aussi* 'también' entre pronombre y verbo como al uso frecuente del pronombre indefinido *ça* (y sus variantes) en función de sujeto.

[13] En una comparación similar (con datos de fuentes menos homogéneas que las usadas aquí), Marins & Soares da Silva (2012) llegan a la conclusión opuesta, según la cual el «sujeto cero» es más frecuente en italiano que en el español peninsular.

3.1. El desarrollo hacia la concordancia sintáctica

El inicio del desarrollo hacia la concordancia sintáctica se debe a dos cambios en el sistema referencial del portugués brasileño. El primero es la sustitución de la segunda persona del singular (2sg), forma de tratamiento informal con *tu*, por la tercera persona del singular (3sg) con la forma de tratamiento *você*, cuyo origen es una alocución formal. Este proceso probablemente comenzó a finales del siglo XVIII y culminó en la primera parte del siglo XX, cuando, excepto en el extremo sur de Brasil, el uso de *tu* y de la 2sg había desaparecido por completo (Martelotta & Cesario 2012: 731-732)[14]. El segundo cambio es la introducción del sintagma nominal referencial *a gente* en función de pronombre para la referencia a la primera persona del plural, que se dio, según Martelotta & Cesario (2012: 733), alrededor de 1970. El pronombre innovador coexiste con la expresión tradicional, como se puede apreciar en el siguiente ejemplo:

(14) *tivemos que ficar esperando esperan/... esperando um mo/ uma hora e meia mais ou menos... até que a enfermeira chamou a gente a gente entrou... tivemos que esperar DE NOVO lá dentro* (Iboruna 74 media 26 fem)
'tuvimos que quedarnos esperan/ ... esperando un mo/ una hora y media más o menos ... hasta que la enfermera nos llamó entramos ... tuvimos que volver a esperar allí dentro'

A diferencia del primer cambio, el uso de *a gente* congruente con la 3sg para la referencia a la primera persona del plural no es específico del PB, sino que se da también, aunque con una frecuencia considerablemente menor, en el PE[15].

Lo que nos interesa aquí es el efecto combinado de estos dos cambios sobre el paradigma verbal del PB, que se compara con los paradigmas del español y portugués europeos en la tabla 1[16].

[14] Sin embargo existen variedades regionales en el norte de Brasil que conservan el uso de *tu* con concordancia de 2sg e incluso 3sg (Cibele Naidhig de Souza, comunicación personal).

[15] En un fragmento elegido al azar de C-Oral-Brasil de 142 181 palabras encontré 563 casos de *a gente* como sujeto pronominal (3,95 casos por 1 000 palabras) mientras que en C-Oral-Rom de Portugal (137 243 palabras) había solo 158 casos análogos de *a gente* (0,91 casos por 1 000 palabras). El recuento no incluye los 32 casos de *a gente* con concordancia de 1pl del PE, fenómeno muy poco frecuente en el PB.

[16] La idea de esta tabla es indicar la referencia de las formas verbales sin apoyo de pronombres, cuyo uso obviamente vacila. Así, por ejemplo, [+H] significa «referencia al hablante» y [+O] «referencia al oyente». [–H, –O] representa la referencia a otra persona o a una entidad concreta o abstracta. La combinación de [+O] con [+formal] representa el tratamiento de *usted* en español, y en portugués

		ESPAÑOL		PORTUGUÉS		
		EUROPEO			EUROPEO	BRASILEÑO
num	**pers**	forma	referencia	forma	referencia	referencia
sg	**1**	*hablo*	[+H]	*falo*	[+H]	[+H]
	2	*hablas*	[+O, –formal]	*falas*	[+O, –formal]	–
	3	*habla*	[–H, –O] [+O, +formal]	*fala*	[–H, –O] [+O, +formal] ([+H, ±O])	[–H, –O] [+O, +formal] [+O, –formal] [+H, ±O]
pl	**1**	*hablamos*	[+H, ±O]	*falamos*	[+H, ±O]	[+H, ±O]
	2	*habláis*	[+O, –formal]	–	–	–
	3	*hablan*	[–H, –O] [+O, +formal]	*falam*	[–H, –O] [+O, +formal] [+O, –formal]	[–H, –O] [+O, +formal] [+O, –formal]

TABLA 1. Paradigmas verbales: español y portugués europeos y portugués brasileño[17]

Lo que salta a la vista en la tabla 1 es la sobrecarga de la 3sg en el PB: además de su función propia, se emplea para el tratamiento tanto formal como informal y para la referencia a 'nosotros'. La consecuencia lógica de esta acumulación de valores es el uso del sujeto pronominal para la desambiguación; efectivamente, Duarte (1995: 19) muestra que en la primera mitad del siglo XX, es decir el período que más o menos coincide con la pérdida definitiva de la 2sg, comienza el incremento del empleo del sujeto pronominal, que luego se va generalizando, es decir que se usa sin que tenga ningún tipo de funcionalidad. Considérense los siguientes ejemplos del uso del pronombre personal de la 3sg del PB de comienzos del siglo XXI:

(15) *assim e a porteira abriu sozinha... [...] e ele já ficou meio assim... aí ele passou... a hora que ele olhou pra trás... tinha uma mula-sem-cabeça... [...] como ele num tinha medo... ele falou assim que a hora que ele olhou... ele arrepiou... ele falou assim que o cabelo arrepiou assim ((mostrando))*
(Iboruna 114 superior 47 fem)
'así el portal se abrió por sí solo ... [...] y él ya quedó medio así ... entonces pasó ... y cuando miró hacia atrás... había una mula sin cabeza[18]... [...] como él no

el de *o senhor* o *a senhora* (o un nombre propio u otro sintagma nominal que indique la función social del oyente en relación con el hablante).

[17] El paradigma del PE está menos diferenciado que el del español, lo cual puede explicar que el uso del pronombre de sujeto exceda el del español (véase cuadro 3). La frecuencia relativamente baja del uso de *a gente* con 3sg en el PE está marcada tipográficamente con paréntesis.

[18] Según una leyenda popular, las concubinas del padre se vuelven mulas después de morir; 'mula-sin-cabeza' es un apodo popular al respecto.

tenía miedo … dijo que en el momento que la vio … tuvo un escalofrío … dijo que el pelo se le erizó así ((mostrando))'

(16) *aí o que aconteceu?… ela veio e falou que ela iria tirar essa criança… entendeu?… que ela i/ tava pensando de abortar essa criança…* (Iboruna 72 primaria 28 fem)
'entonces ¿qué pasó? … ella vino y dijo que iba a quitarse ese niño … ¿entiendes? … que estaba pensando en abortar ese niño …'

Estos dos ejemplos demuestran que el uso del sujeto pronominal no depende de motivaciones ni pragmáticas ni contextuales, que caracterizan su empleo en el español peninsular, y tampoco de la necesidad de desambiguación, sino que parece que se está convirtiendo en obligatorio. Otro indicio de la generalización del pronombre de sujeto es su aparición frecuente con referentes concretos no animados (p. ej. Duarte & Varejão 2003: 106), lo cual no suele ocurrir en español[19].

(17) *e coloco carne moída… e vou fritando ela ai… até ela mudar de cor para ficar dourada… assim que ela começa a dourar…* (Iboruna 80 medio 29 fem)[20]
'pongo carne picada… la voy friendo… hasta que cambie de color para quedar dorada… cuando empiece a dorarse…'

La generalización de los pronombres personales para la referencia inanimada implica su empobrecimiento semántico, puesto que el género gramatical del pronombre ya no es expresión del sexo del referente, sino una simple copia del género gramatical del sustantivo correspondiente. En otras palabras, en el uso con referentes inanimados el pronombre llega a ser producto exclusivo del Nivel Morfosintáctico sin mantener ninguna relación con el Nivel Representacional, que es responsable de introducir la semántica de género que se corresponde con los hechos extralingüísticos (Keizer 2012). Esto significa que el pronombre personal sujeto de la 3sg se está gramaticalizando, en el sentido de desemantizarse y generalizarse (*cfr*. p. ej. Heine &

[19] Aunque no cabe duda de que existen —sobre todo en la lengua escrita formal— casos de sujeto pronominal con referencia a entidades inanimadas abstractas (pero nunca concretas), Fernández Soriano (1999: 1220) afirma: «Los pronombres de sujeto *él/ella*, *ellos/ellas* […] presentan, además, la particularidad de que deben referirse obligatoriamente a personas» (véase también Luján 1999: 1294). Incluso en el español en contacto con el inglés el pronombre sujeto no se suele usar con referencia inanimada. Según Otheguy (2014: 380), el 90% de los sujetos pronominales en el español de Nueva York tiene referencia humana.

[20] En el primer caso no subrayado *ela* tiene función de objeto. Se perdió el clítico de tercera persona en el PB en casi todas sus funciones, conservándose solo el clítico reflexivo en función de paciente (Cyrino 2003).

Narrog 2009: 405-406), aunque es importante notar que está aún lejos del «modelo francés» (véase Kaiser 2009 para un estudio detallado desde un enfoque generativo y Oliveira 2018 para un estudio funcional).

3.2. El desarrollo del marcador de tópico innovador

A partir de las primeras grabaciones del PB oral informal[21], se han registrado casos de «dislocación a la izquierda», tales como los siguientes ejemplos del corpus Iboruna, (18) con referente específico y (19) con referente genérico:

(18) *meu colega… ele começou a tomar à tarde e misturar… éh… cerveja… whisky…*
(Iboruna 81 superior 28 masc)
'mi colega … empezó a beber por la tarde mezclando … cerveza … whisky …'

(19) *às vezes cê deixa a criança numa escoli:nha ou deixa com a emprega:da… a criança… ela vai aprender o quê?..* (Iboruna 116 superior 36 fem)
'a veces dejas el niño en una guardería o lo dejas con la asistenta… el niño … ¿va a aprender qué?

Como ya se ha observado en la sección 2, los sintagmas nominales extrapuestos, tales como *meu colega* en (18) y *a criança* en (19) son actos discursivos dependientes de los principales *ele começou […] whisky* en (18) y *ela vai aprender o quê?* en (19). El que los actos discursivos dependientes no formen parte de los principales se ve claramente en el ejemplo (19), en el cual *a criança* no forma parte de la ilocución interrogativa del acto discursivo que sigue.

Los actos discursivos dependientes con función de orientación tienen una tendencia a irse integrando en el acto discursivo principal hasta llegar a ser tópicos (Hengeveld & Mackenzie 2008: 57, 95-96)[22], lo cual es exactamente lo que ocurre en el portugués de Brasil:

[21] Me refiero a las primeras grabaciones de los proyectos NURC «Projeto Norma Linguística Urbana Culta, Amostra Década 70», que datan de 1970-71, y PEUL «Programa de Estudos sobre o Uso da Língua, Amostra Censo 1980», (con hablantes analfabetos o con educación rudimentaria) de 1970-76. Aunque se trata de una construcción estigmatizada (Maria Luiza Braga, comunicación personal), se encuentran 38 casos en mi fragmento de NURC (497 280 palabras) y solamente 18 en el de PEUL (396 567 palabras).

[22] En los ejemplos ofrecidos por estos autores se trata de SN léxicos marcados morfológicamente. El caso de los pronombres personales de sujeto se discutirá con más detalle en la sección 3.3 de este trabajo.

(20) *Doc.: a relação sua com seu pai hoje é uma relação boa assim:?*
Inf.: é éh meu pai... [...] meu pai ele é uma pessoa assim [...] (Iboruna 72 primaria 28 fem)
'Doc.: tu relación con tu padre hoy ¿es una relación buena?
Inf.: sí ... mi padre ... [...] mi padre es una persona así [...]'

Aparte de la pausa al inicio, la respuesta de (20) forma un solo acto discursivo que, como tal, se corresponde con una frase entonativa única[23]. Este análisis queda confirmado por un estudio acústico de casos análogos en un corpus oral de Rio de Janeiro (Cunha Vieira 2014). De hecho, las construcciones de este tipo forman la gran mayoría (161 casos) de las 212 construcciones de los sintagmas nominales seguidos de *ele* y *ela*, que encontré en el corpus Iboruna[24]. Casi la mitad (78 casos) de los ejemplos con el sintagma nominal integrado tienen referencia inanimada. Lo que más frecuentemente se da son las entidades concretas, como *a costela* en (21), y los estados de cosas, como *a leitura* en (22). Casi siempre se trata de referencias no específicas sino más o menos genéricas, lo cual puede deberse (parcialmente) al género de la entrevista sociolingüística[25].

(21) *a costela ela solta gordura então cê num pode pôr muito óleo né?...*
(Iboruna 70 primaria 27 fem)
'la chuleta suelta grasa entonces no puedes poner mucho aceite, ¿verdad? ...

(22) *Doc.: na sua opinião qual a importân:cia da leitura na alfabetização dos alunos?*
Inf.: a leitura ela é fundamental na alfabetização dos alunos... (Iboruna 88 superior 28 fem)
'Doc.: en tu opiñón ¿cuál es el papel de la lectura en la alfabetización de los alumnos?
Inf.: la lectura es fundamental en la alfabetización de los alumnos ...'

Obviamente el pronombre de sujeto es redundante desde el punto de vista sintáctico en estos casos, y sobrevive en esta posición porque se ha llegado a interpretar como una marca de tópico (Duarte 1995 100-124; Barbosa, Duarte & Kato 2005). La redundancia sintáctica coincide con una

[23] Se podría dudar sobre la adecuación de la representación gráfica de las pausas o de la falta de estas. Del control de algunos de los audios correspondientes a los pronombres en aposición inmediata al SN de sujeto, las transcripciones resultaron ser fiables (véanse también los comentarios al respecto en Wall 2017: 167).

[24] De los 51 restantes había 48 con una pausa entre el sintagma nominal y el pronombre y tres casos que no pude interpretar.

[25] Véase Recalde & Vázquez Rozas (2009) para más detalle. Estas autoras llaman la atención sobre la «artificiosidad del contexto» (2009: 57) de la entrevista sociolingüística.

pérdida completa de la funcionalidad original del pronombre: en la sección 3.1 ya vimos que la aplicación a referentes inanimados implica una pérdida semántica. Cuando se da en función de tópico gana una función pragmática, es decir que la pérdida semántica es precondición para que continúe la gramaticalización o bien que se inicie el proceso de la pragmaticalización[26] del pronombre de sujeto.

3.3. Un posible contraargumento

Hasta ahora solamente nos hemos ocupado de la posición del pronombre personal de tercera persona del singular con respecto al sintagma de sujeto. En esta sección se discutirá el sintagma que sigue a este pronombre, que suele ser el verbo finito, como se puede apreciar en los ejemplos (15)-(19). El fenómeno de presencia de un pronombre duplicado delante del verbo ya lo estudió el neogramático Hermann Paul (1920: 310-311): un pronombre de sujeto duplicado pre o posverbal, cuya función original es algún tipo de énfasis, se puede reducir gradualmente a un clítico y luego a un afijo con función de concordancia verbal, proceso que se da cuando la concordancia verbal preexistente resulta insuficientemente distintiva[27]. La elaboración más conocida de esta idea es la de Givón (1976), quien presenta evidencia de varias lenguas en las que los pronombres de sujeto duplicados preverbales empiezan por convertirse en clíticos y luego se prefijan a la forma verbal hasta llegar a expresar la concordancia con el sujeto. Esquemáticamente, este proceso se puede ilustrar a partir de un simple ejemplo del portugués brasileño *a criança brinca* 'el niño juega', cuyas últimas dos variantes, marcadas por un asterisco, son de naturaleza hipotética:

(23) a. *a criança … ela brinca* [pronombre de sujeto]
b. *a criança ela brinca* [pronombre redundante]
c. **a criança ela-brinca* [clitización del pronombre]
d. **a criança elabrinca* [prefijación del pronombre]

Los casos más convincentes presentados por Givón (1976) son los de las lenguas criollas, en las cuales el origen pronominal de la concordancia verbal

[26] Un caso representativo de pragmaticalización es el desarrollo de partículas pragmáticas a partir de conjunciones (*cfr*. p. ej. Mihatsch 2010 sobre *como* en español).

[27] En el original el contenido resumido aquí reza así: «Das doppelte Ausdrücken des Pronomens tritt ursprünglich nur ein, wo dasselbe besonders hervorgehoben werden soll. […] Die enklitisch angelehnten Pronomina sind mit dem Verbum verschmolzen und haben ihre Subjektsnatur mehr oder weniger eingebüsst. […] Die Hauptursache, welche dazu geführt hat, ist die, dass die Suffixe zur Charakterisierung der Formen nicht mehr ausreichten.» (Paul 1920: 311).

se identifica con facilidad. Givón aclara, además, que no se trata solamente de concordancia de sujeto, sino también de objeto directo o indirecto, a condición de que el origen de las formas correspondientes sean pronombres o clíticos duplicados que precedan o sigan al verbo. En otros estudios posteriores al de Givón se ofrece evidencia, parcialmente hipotética, de más lenguas (p. ej. Lambrecht 1981 sobre el francés popular, y de Groot & Limburg 1986 sobre el abkhaz y el húngaro). Heine & Kuteva (2002) concluyen en su manual de gramaticalización que «[t]he evidence available suggests in fact that third person singular pronouns are the most common source for verbal subject agreement markers» (Heine & Kuteva 2002: 235; véase también Lehmann 2015: 43-45).

El caso del portugués brasileño presentado en este trabajo puede llegar a ser un caso representativo de tal desarrollo: como demostré en la sección 3.1 (tabla 1), el paradigma verbal del PB es bastante reducido y, además, el orden de palabras del PB es casi exclusivamente SVO, por lo cual el pronombre personal redundante casi siempre precede inmediatamente al verbo finito; así que no sería extraño que, a largo plazo, se convirtiese en un marcador de concordancia verbal.

Por tanto, cabe preguntarse si es correcto el análisis presentado en la sección anterior según el cual la integración del pronombre personal dentro del acto discursivo es un proceso de pragmaticalización. En mi opinión, esta tesis se ve respaldada por el hecho de que, sincrónicamente, todavía no existen pruebas del hipotético desarrollo ilustrado en (23): (i) aunque el pronombre personal de sujeto de tercera persona se usa con frecuencia, no es obligatorio, es decir sigue habiendo un cierto grado de congruencia contextual (ver cuadro 3, sección 3), (ii) la construcción en la cual el SN de sujeto está seguido por un pronombre congruente (con o sin pausa) todavía es relativamente infrecuente, y como tal esta construcción está marcada tanto en relación con el SN de sujeto sin pronombre como en relación con el pronombre de sujeto sin SN, y (iii) se supone que, sincrónicamente, el pronombre personal en la posición marcada desempeña una función que, como se comprobará enseguida, es de naturaleza pragmática[28].

En la siguiente sección se describirá en detalle el funcionamiento de esta función pragmática.

[28] Al describir los desarrollos (hipotéticos) mencionados en esta sección, Dik (1997), encuentra pruebas en varias lenguas de que las construcciones del tipo (23b) poseen funciones pragmáticas: «There is evidence from different languages that constructions with Integrated Theme may be exploited for special pragmatic purposes» (Dik 1997: II, 404).

4. El uso del marcador de tópico en el portugués brasileño

El pronombre de sujeto en aposición inmediata puede expresar el tópico en general, el tópico nuevo y el tópico contrastivo (véase sección 2 del presente trabajo). Empecemos por el tópico nuevo:

(24) *quando a polenta tá pronta eu ponho ela numa camada aí eu coloco o molho e uma camada de mussarela... depois ou:tra camada de polenta outra de mussarela... é uma delícia... e esse meu sobrinho ele come até: passar mal* (Iboruna 128 primaria 62 fem)
'cuando la polenta está lista la pongo en una capa luego pongo la salsa y una capa de mozzarella ... después otra capa de polenta otra de mozzarella... es una delicia... este sobrino mío puede comerlo hasta marearse'

Como el contexto de (24) es la descripción de una receta de cocina, el referente *esse meu sobrinho ele* es un tópico nuevo, que en gramática discursivo-funcional se presenta como una combinación de las funciones de tópico y de foco (TopFoc) en este referente. El ejemplo (24) es uno de los pocos casos en mi corpus con un tópico completamente nuevo. Más comunes son los casos en los que el referente es aparentemente nuevo, pero de hecho no lo es:

(25) *Doc.: pra fazer a barreira tem: um determinado número de jogadores [...]*
Inf.: não [...] o goleiro ele pode pedir o tanto que quiser... (Iboruna 119 superior 54 masc)
'Doc.: para hacer una barrera ¿hay un determinado número de jugadores?
Inf.: no [...] *el portero* puede pedir a cuántos quiera ...'

El referente *o goleiro* en la conversación sobre el fútbol es lo que según Dik (1997: I, 324) constituye un subtópico, es decir un referente inferible de la conversación anterior y del conocimiento del mundo (almacenados en el componente contextual). Así que de hecho se trata, en términos discursivo-funcionales, de un simple tópico. Un caso similar es el siguiente:

(26) *Doc.: cê poderia me falar como é que se FAZ alguma coi:sa?*
Inf.: sim poderia... por exemplo... o meu ramo de atividade hoje é programação... como eu sou analista de sisTEma... então eu desenvolvo sistemas principalmente pra área comercial administrativa... éh:... um programa ele é uma coisa muito interessante...
(Iboruna 99 primaria 36 masc)
'Doc.: ¿me podría hablar de cómo es que hace alguna cosa?
Inf.: sí claro ... por ejemplo ... mi campo de actividad ahora es la programación ... ya que soy analista de sistemas ... entonces yo desarrollo sistemas principalmente para el área comercial administrativa ... un programa es una cosa muy interesante ...'

Aquí también, el referente de *um programa*, a pesar de no haber sido mencionado antes, no es realmente nuevo ya que es fácil de inferir sobre la base del contexto inmediatamente anterior. Lo interesante de este ejemplo es el artículo indeterminado, que indica que el referente no es específico ni para el hablante ni para el oyente. No obstante, se puede considerar como «dado» en el sentido de ser un referente tanto «activo» en la conversación como «compartido» por los interlocutores (véase García Velasco 2018 para más detalles).

En el caso siguiente se da exactamente lo opuesto: aunque el documentalista [Doc.] ya ha introducido una entidad genérica con la propiedad *perro*, su interlocutor [Inf.] lo corrige notificándole que su perro específico es distinto de lo que parece presuponer el otro:

(27) *Doc.: e o cachorro dorme lá do lado [da cama]?*
Inf.: não: *o meu cachorro ele é grande ele dorme lá fora...* (Iboruna 28 media 23 masc)
'Doc.: ¿y el perro duerme allí al lado [de la cama]?
Inf.: no mi perro es grande y duerme allí fuera ...'

Esta «corrección» por parte del informante se marca por la combinación del artículo con el pronombre posesivo *o meu*, mientras que lo no marcado sería *meu cachorro*. Por tanto, el ejemplo (27) representa un caso de tópico dado que a la vez es contrastivo[29].

Son escasos en mi corpus los tópicos claramente contrastivos, dos de los cuales son los ejemplos (28) y (29):

(28) *Doc.: e como que joga?*
Inf.: os peões... andam... em diagonal... e só pode come:r de frente... aí eles andam em diagonal... e come de frente... a torre só anda re:to e vira... num pode andar diagonal... o cavalo ele anda:... éh duas casas e vira... duas casas pra frente duas casa pro lado esquerdo e duas casas pro lado... direito ou pa trás... (Iboruna 6 primaria 10 fem)
'Doc.: ¿y cómo es que se juega?
Inf.: los peones andan... en diagonal ... y solo pueden comer de frente... entonces andan en diagonal y come [*sic*] de frente... la torre solo anda recto y gira... no puede andar diagonal ... el caballo anda... eh dos casillas y gira... dos casillas de frente dos casillas hacia el lado izquierdo y dos casillas hacia el lado ... derecho o para atrás ...'

[29] Agradezco a Lachlan Mackenzie el haberme llamado la atención sobre este caso.

(29) *Doc.: [...] como é que cê: corta o cabelo de um homem por exemplo?*
Inf.: [...] cabelo de homem é fácil assim... éh porque o cabelo de mulher tem que dividir TOdas as mecha do cabelo entendeu?... agora homem não... homem você corta as laterais... e depois cê mexe a parte de cima... dependendo o o que for feito no cabelo o corte de cabelo que você quer fazer... faz o pezi:nho entendeu? do cabelo... em volta da ore:lha... porque geralmente a mulher ela num tem esse pezinho que o homem tem... (Iboruna 72 primaria 28 fem)
'Doc.: [...] ¿cómo es que cortas el pelo de un hombre por ejemplo?
Inf.: [...] el pelo de hombre es fácil ... eh porque el pelo de mujer tiene que dividirse todo en mechas, ¿no? ... ahora los hombres no ... con los hombres cortas los lados ... y después vas por la parte de encima ... dependiendo de lo que se vaya a hacer con el pelo o del corte de pelo que tú quieras ... haces la patilla, ¿no? del pelo ... alrededor de la oreja ... porque generalmente las mujeres no tienen esa patilla que los hombres tienen ...'

En el ejemplo (28), sobre las reglas del ajedrez, se contrasta la pieza del caballo con el peón y la torre por tener más posibilidades de movimiento dentro del juego, es decir que el referente *cavalo* se marca como tópico contrastivo en relación a *os peões* y *a torre*. En el ejemplo (29) se habla de lo que la peluquera puede hacer con el pelo del hombre y la hablante contrasta el hecho de que el hombre tiene cabello delante de la oreja que la mujer no tiene. Así que *a mulher* se marca como tópico contrastivo.

Los casos mayoritarios y más interesantes son los del tópico general. Los siguientes ejemplos son representativos:

(30) [sobre las elecciones a la alcaldía]
agora pro fim... que ficou agora pro segundo turno o Antu/ aqui pra prefeito o Antunes... e o: Edinho... agora eu acho que o Antunes ele já tá meio cansado pelo que eu ouvi hoje uma entrevista dele... (Iboruna 152 superior 76 fem)
'ahora al final... que quedó ahora para la segunda vuelta el Antu/ aquí para alcalde el Antunes o el Edinho ... ahora yo creo que el Antunes ya está medio cansado por lo que vi hoy en una entrevista con él'

(31) *tem uma escola aqui no centro que é boa que é o Anísio e tem o:: Edmur também... e: o ônibus... tem um ônibus né? mas só que o ônibus ele tá precário... tá precário mesmo*
(Iboruna 74 media 26 fem)
'hay una escuela aquí en el centro que es buena que es el Anísio y está el Edmur también ... y el autobús ... hay un autobús ¿eh? pero solo es que el autobús es desastre ... es un verdadero desastre'

(32) *em primeiro lugar cê tem que fazer o creme né? porque o creme ele é que/ ele vai ficar quente né? então cê... faz o creme... enquanto ele esfria você faz o bolo porque o bolo é rapidinho né?* (Iboruna 89 primaria 42 masc)
'en primer lugar tienes que hacer la crema ¿no? porque la crema es que/ va a estar caliente, ¿no? entonces ... haces la crema ... mientras se enfría haces el pastel porque el pastel es rapidito, ¿no?'

En todos estos casos el referente marcado ya se ha mencionado una o más veces en el contexto que precede, es decir que no es nuevo, sino que forma parte del conocimiento compartido de los interlocutores. La marca de tópico está motivada por el hecho de que la/el hablante tematice el referente para reconsiderarlo desde una perspectiva subjetiva, introducida típicamente por *eu acho que* 'yo creo que', ilustrado en (30), o para añadir una explicación, frecuentemente introducida por *só que* 'solo es que' en (31), o por *porque* en (32). Estos casos son sintomáticos del uso de la marca innovadora, pues esta se emplea para indicar algo como «ahora voy a hablar más sobre este tema», lo cual es característico de la función de una marca explícita de tópico.

5. La expresión de tópico y foco en el español europeo, el PE y el PB

En la introducción a este trabajo he mencionado someramente las posibilidades del español y del portugués para expresar las funciones de tópico y foco en cláusulas transitivas. En esta sección los datos descritos en las secciones 3 y 4 se van a comparar con los del español y portugués europeos, con objeto de obtener una comparación sistemática de estas tres variedades en cuanto a la expresión morfosintáctica de las dos funciones pragmáticas citadas con referentes en función sintáctica de sujeto.

Antes de entrar en detalles, hace falta mencionar la construcción de tópico generalmente llamada «elevación» (*raising*) común a varias lenguas (Dik 1997: II, 344-351), entre ellas el portugués (Dik 1981; Gonçalves 2004). Esta construcción se registra, sin embargo, con un número muy limitado de verbos, típicamente con *parecer*, ilustrado en (33) con un ejemplo adaptado de Hernanz (1999: 2229-2230):

(33) a. Parece que las niñas tienen razón.
b. Las niñas parece que tienen razón.

El sujeto de la cláusula completiva *las niñas* de (33a) aparece en (33b) en posición inicial. Según la NGLE (2009: 2834) el efecto es que el referente

«elevado» adquiere función de tópico[30]. Dada la aplicación restringida de esta construcción (y sus múltiples variantes) ya no se discutirá en el presente contexto.

El resto de esta sección se dedicará a las demás construcciones de foco y tópico. Empezaré por revisar las construcciones del español para luego pasar a las de las dos variedades del portugués. Con respecto al español peninsular, se ha visto que el foco con referentes de sujeto se expresa más fácilmente por la posición posverbal de este:

(34) todo eso lo hizo mi marido (AdH 35 media 67 fem)

Alternativamente, la función de foco se puede expresar por medio de una construcción pseudohendida. En el ejemplo (35) se hace referencia a un estado de cosas focal:

(35) luego ya me dediqué porque lo que me gustaba era:/ esto/ la autoescuela más que nada el trato con la gente/lo que me gusta es:/ el contacto personal (AdH 25 media 35 masc)

Como se indica en la introducción a este trabajo, la función de tópico se puede expresar por medio del orden de constituyentes solamente cuando el referente en cuestión no tiene función de sujeto. En principio, la expresión del tópico también es posible con referentes de sujeto, por medio de la construcción hendida, pero es un fenómeno de la lengua escrita:

(36) De hecho, la víspera, tras conocer el dramático desenlace del secuestro, Ardanza ha hecho unas durísimas declaraciones acusando directamente a Herri Batasuna de complicidad con ETA y con sus atentados:
– ETA es quien ha dado cumplimiento a su amenaza pero ETA no es la única responsable de lo ocurrido. (CREA, Carmen Gurruchaga e Isabel San Sebastián, *El árbol y las nueces: la relación secreta entre ETA y PNV*. Madrid: Temas de Hoy, 2000)

En el portugués europeo el foco en las cláusulas transitivas raramente se expresa al final de la cláusula (*cfr*. Mackenzie 2008: 70-71). Lo común es que la función pragmática de foco del referente del sujeto se exprese por medio de una construcción pseudohendida:

[30] El término «elevación» es de corte generativo e implica una transformación de movimiento, que no existe en la teoría discursivo-funcional, como ya se ha mencionado en la sección 2. Una explicación discursivo-funcional de este fenómeno se encuentra en García Velasco (2013).

(37) *e o mais engraçado é que quem nos assaltou o carro foi o gang do multibanco*
(COR-P, conversación libre, «Conciertos» media 26-40 masc)
'y lo más gracioso es que quien nos asaltó el coche era la pandilla del cajero automático'

Para la expresión de tópico se puede recurrir, igual que en español, a la construcción hendida. En (38) el referente *a vítima* está marcado como tópico por medio de la construcción hendida. Aunque no se ha usado el lexema *vítima* anteriormente, está claro que hace referencia a la mujer mencionada en el contexto inmediatamente anterior.

(38) *O homem de 26 anos acusado de homicídio qualificado por ter degolado a mulher, disse esta quinta-feira em tribunal que não tinha intenção de matar, acrescentando que foi a vítima quem primeiro agarrou em a faca e o agrediu.*
(CdP-Portugal, prensa, 24-05-2012)
'El hombre de 26 años acusado de homicidio cualificado por haber degollado a una mujer, dijo este jueves en el tribunal que no había tenido la intención de matarla, añadiendo que fue la víctima quien primero cogió la navaja y lo agredió.'

En el portugués de Brasil, finalmente, la función de foco se suele expresar por medio de una pseudohendida, al igual que en el portugués europeo:

(39) *essa história quem me contou foi meu vô* (Iboruna 27 primaria 25 masc)
'esta historia quien me la contó fue mi abuelo'

Con respecto a la función de tópico, demostré en las secciones 3 y 4 que se expresa por medio de la construcción innovadora, de la cual (40) es otro ejemplo:

(40) *pra cada copo de arroz... eu costumo dobrar o copo de água... só que tem uma coisa o arroz cê num pode lavar... o arroz e deixar o arroz quietinho lá... não o arroz ele tem que ser lavado na ho::ra que você for começar fazer o arroz...*
(Iboruna 88 superior 28 fem)
'para cada taza de arroz ... yo suelo poner dos de agua ... solo es que el arroz no puedes lavar ... el arroz y dejar el arroz allí tranquilamente ... no el arroz se tiene que lavar en el momento que empiezas a hacer el arroz'

El siguiente ejemplo ilustra la expresión del tópico por la construcción hendida.

(41) *O compositor, cantor e baixista Sting, vai muito bem em sua carreira solo. Dos três integrantes do extinto conjunto 'The Police', ele foi quem se saiu melhor.*
(CdP-Brasil, «Police ainda é referência no rock». 1997)
'El compositor, cantante y bajista Sting, va muy bien en su carrera en solitario. De los tres integrantes del antiguo conjunto 'The Police', a él es a quien le ha ido mejor.'

El texto de este ejemplo versa sobre el artista Sting, y la mención pronominal *ele* está marcada como tópico por la construcción hendida, ya que la marca de tópico innovadora solo se puede aplicar a sintagmas nominales léxicos y no a los pronombres personales.

La tabla 2 resume las realizaciones de tópico y foco de los referentes con función sintáctica de sujeto en las variedades.

	Español europeo	Portugués europeo	Portugués brasileño
Foco	marca específica pseudohendida	pseudohendida	pseudohendida
Tópico	hendida	hendida	marca específica hendida

Tabla 2. Realizaciones de foco y de tópico en el español y portugués europeos y el portugués de Brasil

La diferencia entre las construcciones hendida y pseudohendida y las expresiones de foco del español y de tópico en el PB está en que las primeras son expresiones que el español y el portugués tienen en común con las lenguas románicas, germánicas y muchas otras más (*cfr*. Dik 1997: II, 291-312). Esto significa que las construcciones hendidas en general no son específicas de las lenguas iberorrománicas, por lo que no tienen interés en el contexto de este trabajo. Lo que es relevante aquí son las formas de expresar las funciones pragmáticas específicas en el español europeo y el portugués brasileño, el foco en español y el tópico en el PB.

6. Conclusión

En este capítulo he comparado el portugués de Brasil con el portugués y el español europeos mostrando que ninguna variedad del portugués posee una realización específica de índole morfosintáctica para la función de foco y que el portugués y el español europeos carecen de una expresión morfosintáctica de tópico específica. Esto no implica que no se puedan codificar estas funciones: existen construcciones hendidas y pseudohendidas que sirven para su

expresión morfosintáctica, procedimientos compartidos con una gran variedad de lenguas, entre ellas las germánicas.

La situación es distinta en español, que marca el foco del referente del sujeto por la posición posverbal, y en el portugués de Brasil con su expresión innovadora de la función de tópico.

Con respecto al portugués de Brasil, he descrito cómo se ha llegado a generalizar el uso del pronombre de tercera persona (inicialmente) debido a la sobrecarga de la concordancia de 3sg, que a lo largo del siglo XX llegó a usarse para la alocución informal y para la referencia a la primera persona del plural. La generalización del uso del pronombre de sujeto también implica su empleo sistemático con entidades inanimadas. Esto significa que el género del pronombre ha dejado de servir exclusivamente como expresión de una propiedad semántica (el sexo del referente) para llegar a ser una simple copia del género gramatical del sustantivo nuclear del sintagma nominal correspondiente. Esta desemantización ha hecho posible su pragmaticalización —en esta construcción específica— de tal modo que, con independencia de la animación del referente, ha llegado a perder por completo su función de pronombre para convertirse en una marca de tópico. Es decir, la pérdida semántica es precondición para que continúe el proceso de gramaticalización hacia una función nueva, pragmática, o bien interpersonal.

A consecuencia del proceso de pragmaticalización, todavía incipiente, el portugués brasileño se está convirtiendo en una lengua que marca el tópico, lo que contrasta con el español, que es una lengua que marca el foco. Así que dos lenguas tan relacionadas genéticamente han desarrollado propiedades opuestas en el dominio de las funciones pragmáticas.

Abreviaturas

2sg = segunda persona de singular
3sg = tercera persona de singular
A = acto discursivo
C = contenido comunicado
GDF = gramática discursivo-funcional
H = hablante
IP = frase entonacional (*Intonational Phrase)*
M = movimiento (≈turno en la interacción oral)
n = cualquier número
φ = cualquier función pragmática

O = oyente
P = participante
PB = portugués brasileño
PE = portugués europeo
R = referente
T = adscripción

Corpus

AdH: Moreno Fernández, Francisco, Ana M. Cestero Mancera, Isabel Molina Martos & Florentino Paredes García (eds.) (2002-07): *La lengua hablada en Alcalá de Henares*. Alcalá: Universidad de Alcalá.

CdP: Davies, Mark (online) *Corpus do Português*. https://www.corpusdoportugues.org/

COB: Raso, Tommaso & Heliana Mello (eds.) (2012): *O corpus C-ORAL-BRASIL*. Belo Horizonte: Editora da Universidade Federal de Minas Gerais.

COR-P: Cresti, Emanuela & Massimo Moneglia (eds.) (2005): *C-ORAL-ROM: Integrated reference corpora for spoken Romance languages* – Portuguese. Amsterdam: John Benjamins.

CREA: Real Academia Española (en línea). *Corpus de referencia del español actual* (CREA). http://www.rae.es

CSF: Beeching, Kate: *Corpus of Spoken French* [1980-90]. Bristol: University of West England. https://www1.uwe.ac.uk/cahe/research/bristolcentreforlinguistics/researchatbcl/iclru.aspx

Iboruna: Gonçalves, Sebastião Carlos Leite (coord.): *Banco de dados Iboruna: amostras eletrônicas do português falado no interior paulista* [2002-07]. http://www.alip.ibilce.unesp.br/iboruna

NURC: *Projeto Norma Linguística Urbana Culta* - RJ, Amostra Década de 70, Facultade de Letras, Universidade Federal de Rio de Janeiro (UFRJ). http://www.nurcrj.letras.ufrj.br/

PEUL: *Programa de Estudos sobre o Uso da Língua*, Amostra Censo 1980, Facultade de Letras, Universidade Federal de Rio de Janeiro (UFRJ).

PF-P: Centro de Linguística da Universidade de Lisboa (CLUL): *Português Falado* - Portugal [1990-1999]. http://www.clul.ulisboa.pt/

Referencias bibliográficas

Andrade Peres, João & Telmo Móia (1995): *Áreas críticas da língua portuguesa*. Lisboa: Caminho.

Barbosa, Pilar, Maria Eugênia Duarte & Mary Kato (2005): «Null subjects in European and Brazilian Portuguese», *Journal of Portuguese Linguistics* 4, pp. 11-52. https://doi.org/10.5334/jpl.158

Bosque, Ignacio & Javier Gutierrez Rexach (2009): *Fundamentos de sintaxis formal*. Madrid: Akal.

Cunha Vieira, André Felipe (2014): *Construção SNpleno-tópico$_i$ + SNpro$_i$ + verbo no português do Brasil: uma análise funcional baseada no uso*. Tesis de máster. Universidade Federal de Rio de Janeiro.

Cyrino, Sonia Maria Lazzarini (2003): «Para a história do português brasileiro: a presença do objeto nulo e a ausência dos clíticos», *Letras de Hoje* 38/1, pp. 31-47.

de Groot, Casper & Machiel Limburg (1986): «Pronominal elements: diachrony, typology, and formalization in Functional Grammar», *Working Papers in Functional Grammar* 12, pp. 1-70.

Dik, Simon C. (1981): «The interaction of subject and topic in Portuguese», in Machtelt A. Bolkestein, Henk Combé, Simon C. Dik, Casper de Groot, Jadranka Gvozdanović, Albert Rijksbaron & Co Vet: *Predication and expression in Functional Grammar*. London: Academic Press, pp. 165-184.

Dik, Simon C. (1997): *The theory of Functional Grammar*, ed. por Kees Hengeveld, 2 vols. Berlin: Mouton de Gruyter.

Duarte, Maria Eugênia (1995): *A perda do princípio «Evite pronome» no português brasileiro*. Tesis de doctorado. Universidade de Campinas, Instituto de Estudos Linguísticos.

Duarte, Maria Eugênia & Filomena Varejão (2013): «Null subjects and agreement marks in European and Brazilian Portuguese», *Journal of Portuguese Linguistics* 12/2, pp. 101-123. https://doi.org/10.5334/jpl.69

Fernández Soriano, Olga (1999): «El pronombre personal; formas y distribuciones; pronombres átonos y tónicos», in Ignacio Bosque & Violeta Demonte (eds.): *Gramática descriptiva de la lengua española*. Madrid: Espasa Calpe, pp. 1209-1273.

Fuentes Rodríguez, Catalina (1997): «Los conectores en la lengua oral: *es que* como introductor del enunciado», *Verba* 24, pp. 237-263.

García Velasco, Daniel (2013): «Raising in Functional Discourse Grammar», in J. Lachlan Mackenzie & Hella Olbertz (eds.): *Casebook in Functional Discourse Grammar*. Amsterdam: John Benjamins, pp. 249-275. https://doi.org/10.1075/slcs.137.10vel

García Velasco, Daniel (2014): «Activation and the relation between context and grammar», *Pragmatics* 24/2, pp. 297-316. https://doi.org/10.1075/prag.24.2.06vel

García Velasco, Daniel (2018): «Cognitive status and referential acts in Functional Discourse Grammar», *Quaderns de Filologia: Estudis Linguistics* 23, pp. 155-175. https://doi.org/10.7203/qf.23.13525

Givón, Talmy (1976): «Topic, Pronoun and grammatical agreement», in Charles N. Li (ed.): *Subject and topic*. New York: Academic Press, pp. 149-188.

Gonçalves, Sebastião Carlos Leite (2004): «Verbo *parecer* no PB: um caso de gramaticalização?» *Sínteses: Revista dos Cursos de Pós-Graduação* 9, pp. 195-209.

Hannay, Mike & Elena Martínez Caro (2008): «Clause-final focus constituents in English and Spanish», in María de los Ángeles Gómez González, J. Lachlan Mackenzie & Elsa González Álvarez (eds.): *Languages and cultures in contrast: new directions in contrastive linguistics*. Amsterdam: John Benjamins, pp. 33-68.

Heine, Bernd & Tania Kuteva (2002): *World lexicon of grammaticalization*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511613463

Heine, Bernd & Heiko Narrog (2009): «Grammaticalization and linguistic analysis», in Bernd Heine & Heiko Narrog (eds.): *The Oxford handbook of linguistic analysis*, pp. 401-424. https://doi.org/10.1093/oxfordhb/9780199544004.013.0016

Hengeveld, Kees (2012): «Referential markers and agreement markers in Functional Discourse Grammar», *Language Sciences* 34, pp. 468-479. https://doi.org/10.1016/j.langsci.2012.03.001

Hengeveld, Kees (2017): «A hierachical approach to grammaticalization», in Kees Hengeveld, Heiko Narrog & Hella Olbertz (eds): *The grammaticalization of tense, aspect, modality and evidentiality. A functional perspective*. Berlin: de Gruyter, pp. 13-43.

Hengeveld, Kees & J. Lachlan Mackenzie (2008): *Functional Discourse Grammar: a typologically-based theory of language structure*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199544004.013.0015

Hengeveld, Kees & J. Lachlan Mackenzie (2011): «La Gramática discursivo-funcional» [traducción española de Daniel García Velasco], *Moenia: Revista lucense de lingüística e literatura* 17, pp. 5-45.

Hernanz, María Lluïsa (1999): «El infinitivo», in Ignacio Bosque & Violeta Demonte (eds.): *Gramática descriptiva de la lengua española*. Madrid: Espasa Calpe, pp. 2196-2356.

Kaiser, Georg A. (2009): «Losing the null subject: a contrastive study of (Brazilian) Portuguese and (Medieval) French», in Georg A. Kaiser & Eva Maria Remberger (eds.): *Proceedings of the Workshop «Null-subjects, expletives,*

and locatives in Romance». FB Sprachwissenschaft, Universität Konstanz, pp. 131-156.

Keizer, Evelien (2012): «English proforms in Functional Discourse Grammar», *Language Sciences* 34, pp. 400-420. https://doi.org/10.1016/j.langsci.2012.02.009

Kroon, Caroline (1995): *Discourse particles in Latin*. Amsterdam: Gieben.

Lambrecht, Knud (1981): *Topic, antitopic and verb agreement in non-standard French*. Amsterdam: John Benjamins. https://doi.org/10.1075/pb.ii.6

Lehmann, Christian (2015): *Thoughts on grammaticalization*. Berlin: Language Science Press [3a edición revisada]. https://doi.org/10.26530/OAPEN_603353

Leonetti, Manuel (2011): «La expresión de la estructura informativa en la sintaxis: un parámetro de variación en las lenguas románicas», *Romanistisches Jahrbuch* 61, pp. 338-355.

Luján, Marta (1999): «Expresión y omisión del pronombre personal», in Ignacio Bosque & Violeta Demonte (eds.): *Gramática descriptiva de la lengua española*. Madrid: Espasa Calpe, pp. 1275-1315.

Mackenzie, J. Lachlan (2008): «The contrast between pronoun position in European Portuguese and Castilian Spanish: an application of Functional Grammar», in María de los Ángeles Gómez González, J. Lachlan Mackenzie & Elsa M. González Álvarez (eds.): *Current trends in contrastive linguistics: functional and cognitive perspectives*. Amsterdam: John Benjamins, pp. 51-75. https://doi.org/10.1075/sfsl.60.05mac

Marins, Juliana Espósito & Humberto Soares da Silva (2012): «A representação do sujeito pronominal no grupo românico: espanhol e italiano em contraste com o português», *Caligrama* 17/2, pp. 91-114. https://doi.org/10.17851/2238-3824.17.2.91-114

Martelotta, Mário Eduardo & Maura Cezario (2009): «Grammaticalization in Brazilian Portuguese», in Bernd Heine & Heiko Narrog (eds.): *The Oxford handbook of grammaticalization*. Oxford: Oxford University Press, pp. 729-739. https://doi.org/10.1093/oxfordhb/9780199586783.013.0060

Martínez Caro, Elena (2007): «Pragmatic frames, the thetic-categorical distinction and Spanish constituent order», *Alfa: Revista de Linguística* 51/2, pp. 119-142.

Martins, Ana Maria (2011): «Scrambling and information focus in old and contemporary Portuguese», *Catalan Journal of Linguistics* 10, pp. 133-158. https://doi.org/10.5565/rev/catjl.35

Mihatsch, Wiltrud (2010): «Sincronía y diacronía del aproximador *como*», *Revista Internacional de Lingüística Iberoamericana* 8/2, pp. 175-201.

NGLE = Real Academia Española / Asociación de Academias de la lengua española (2009): *Nueva gramática de la lengua española*. Madrid: Espasa.

Oliveira, Taísa P. (2018): «Subject expression in Brazilian Portuguese», in Evelien Keizer & Hella Olbertz (eds.): *Recent developments in Functional Discourse Grammar*. Amsterdam: John Benjamins, pp. 207-232. https://doi.org/10.1075/slcs.205.07deo

Otheguy, Ricardo (2014): «Remarks on pronominal perseveration and functional explanation», en Andrés Enrique Arias, Manuel J. Gutiérrez, Alazne Landa & Francisco Ocampo (eds.), *Perspectives in the study of Spanish language variation: papers in honor of Carmen Silva-Corvalán*. Santiago de Compostela: Universidade de Santiago de Compostela (*Verba*, Anexo 72), pp. 373-396. http://dx.doi.org/10.15304/va.2014.701

Paul, Hermann (1920): *Prinzipien der Sprachgeschichte*. Halle: Max Niemeyer [5a edición revisada].

Recalde, Montserrat & Victoria Vázquez Rozas (2009): «Problemas metodológicos en la formación de corpus orales», in Pascual Cantos Gómez & Aquilino Sánchez Pérez (eds): *A survey of corpus-based research*. Murcia: Asociación Española de Lingüística del Corpus, pp. 51-64. https://www.um.es/lacell/aelinco/contenido/pdf/4.pdf

Siewierska, Anna (2004): *Person*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511812729

Wall, Albert (2017): *Bare nominals in Brazilian Portuguese: an integral approach*. Amsterdam: John Benjamins. https://doi.org/10.1075/la.245

Zubizarreta, María Luisa (1999): «Las funciones informativas: tema y foco», in Ignacio Bosque & Violeta Demonte (eds.): *Gramática descriptiva de la lengua española*. Madrid: Espasa Calpe, pp. 4214-4246.

CORPUS DE REFERENCIA DO GALEGO ACTUAL (CORGA): COMPOSICIÓN, CODIFICACIÓN, ETIQUETAXE E EXPLOTACIÓN

The Reference corpus of present-day Galician (CORGA): composition, codification, POS-tagging and use

Eva María Domínguez Noya (USC/CIRP)
María Sol López Martínez (USC/CIRP)
Francisco Mario Barcala Rodríguez (NLPgo Technologies S.L.)

Resumo

O Corpus de Referencia do Galego Actual (CORGA), accesible no enderezo http://corpus.cirp.gal/corga, é un corpus documental aberto que abrangue cronoloxicamente dende 1975 ata a actualidade, cuxo obxectivo é fornecer datos para o estudo da lingua galega actual dende múltiples perspectivas: léxica, morfolóxica, sintáctica, fraseolóxica, terminolóxica, comunicativa etc. O corpus, enriquecido automaticamente coa lematización e etiquetaxe morfosintáctica dos seus textos, contén 40 178 271 palabras ortográficas pertencentes maioritariamente a distintos tipos de textos escritos representativos do galego actual, mais tamén inclúe 25 horas de transcricións ortográficas de programas de radio nas que se aliña o texto coa voz.

Neste traballo abórdase primeiramente unha caracterización xeral do deseño e composición do corpus, e logo, no núcleo do artigo, descríbese o seu proceso de creación: a dixitalización, a codificación e maila etiquetaxe dos textos que contén. Os criterios seguidos na creación e as decisións tomadas sobre a etiquetaxe están estreitamente vinculados a un contexto de contacto de linguas no que ademais ten lugar o proceso de normativización do galego. Por último, proporciónase unha minuciosa descrición das posibilidades de recuperación de información que ofrece a aplicación da consulta en liña.

Palabras chave: corpus, galego, codificación, etiquetaxe morfosintáctica, uso

Abstract

The *Corpus de Referencia do Galego Actual* (CORGA), accessible online at http://corpus.cirp.gal/corga, is an open corpus that covers the period from 1975 until the present. Its objective is to supply data for the study of present-day Galician from different perspectives, such as the lexicon, morphology and syntax, as well as terminology, phraseology and communication. The corpus, which has been automatically enriched with the lemmatization and morphosyntactic tagging, consists of 40,178,271 orthographic words, most of which belong to different types of written texts that are representative of present-day Galician. It also includes 25 hours of orthographic transcriptions of radio programmes with speech-to-text alignment.

This chapter begins with a general overview of the design and composition of the corpus. The main part is dedicated to describing the process of its creation: the digitalization, codification and POS-tagging of the corpus texts. It is shown that the criteria guiding the creation of the corpus and the decisions to be taken with respect to morphosyntactic tagging are closely related to the context of language contact, which, again, plays a crucial role in the process of normativization of Galician. In addition, the chapter contains a detailed description of the possibilities of information retrieval the online application offers the users.

Keywords: corpus, Galician, codification, morphosyntactic tagging, use

1. Introdución

O *Corpus de Referencia do Galego Actual* (CORGA) é un proxecto que se desenvolve no Centro Ramón Piñeiro para a investigación en humanidades (CRPIH) desde o ano 1993, data na que, como sinalan Rojo *et al.* (2016: 445), «a lingüística de corpus adquirira xa un altísimo nivel de logros». Como corpus de referencia, é unha colección de documentos que se almacenan en formato electrónico na que se integran distintos tipos de textos representativos do uso da lingua galega actual: xornais, revistas, blogs, ensaios e textos de ficción (novelas, relatos curtos, obras de teatro e guións). Así mesmo, dende a versión 3.0 dáselle cabida ó rexistro oral coa inclusión de transcricións de programas de radio da década dos 90, transcricións que paulatinamente se irán estendendo ós demais períodos cronolóxicos establecidos.

Nestes tempos nos que a capacidade de almacenamento nos servidores é case infinita, os esforzos que se empregan na construción dun corpus coma o CORGA —de tamaño medio, cun deseño previo e unha codificación rigorosa como veremos ó longo deste traballo— repercuten de xeito moi positivo na fiabilidade dos datos que achega o sistema á comunidade científica. Na construción deste corpus incorpóranse documentos que se atopan en Inter-

net, seleccionados pola súa tipoloxía, como son textos xornalísticos, revistas ou blogs; pero tamén, grazas á amabilidade e á boa disposición de autores e editoriais[1], obras literarias e ensaios recentes, naturalmente «coa garantía de que os usuarios do corpus poden recuperar liñas de concordancias, pero nunca a obra completa nin fragmentos representativos dela» (Rojo *et al*. 2016: 447).

Cronoloxicamente, o CORGA abrangue desde o ano 1975 ata a actualidade e os textos que recompila reflicten a lingua real que se emprega na escrita e tamén na oralidade. Como é sabido, as linguas presentan variación diatópica, diacrónica, diastrática e diafásica, pero o galego, ó igual ca outras linguas, presenta tamén variación relacionada coa súa propia historia[2], coa ausencia dunha norma oficial para a escrita e posteriormente coa diversidade gráfica[3] provocada polas diferentes propostas normativas. Son por tanto habituais as interferencias derivadas do contacto de linguas, a variación e a falta de coherencia nas solucións gráficas e morfolóxicas.

Tal como se acaba de sinalar, os primeiros textos do CORGA datan do ano 1975, momento no que o galego non tiña recoñecida a oficialidade nin dispuña dunha proposta normativa para a escrita, inda que desde principios dos anos 70 tanto a RAG coma outras institucións eran conscientes de que o pulo, o crecemento e o prestixio da literatura galega «esixían urxentemente unha codificación da ortografía usual» (RAG 1970: 8)[4]. Simultaneamente, o Instituto da Lingua Galega (ILG) elabora «un método práctico para el aprendizaje y perfeccionamiento del gallego»[5]. Estas propostas non acadan o consenso nin o compromiso para o seu uso[6], pero os cambios políticos que se producen a finais dos anos 70 fan necesario dispor dunha norma unifi-

---

[1] Quede aquí constancia do noso agradecemento ás editoriais que ó longo destes anos nos veñen facilitando os textos en formato pdf para a súa incorporación ó corpus.

[2] A lingua galega pasou ó longo da historia por unha serie de vicisitudes que non favoreceron a normalización social e, consecuentemente, tampouco a normativización. Ata mediados do século XIX non se publican as primeiras gramáticas e dicionarios, así que as persoas que empregan nos seus escritos o galego de maneira xeral reproducen a fala da súa zona.

[3] A primeira proposta normativa data de 1933 e procede do Seminario de Estudos Galegos que na súa sesión do 15 de outubro acordou a publicación de *Algunhas normas pra a unificazón do idioma galego,* que non acadou o consenso necesario.

[4] En 1970 a RAG aproba as *Normas ortográficas do idioma galego*, e no 1971 completa o documento anterior cunhas *«normas morfolóxicas»* que publica co título *Normas ortográficas e morfolóxicas do idioma galego.*

[5] Deste método publicáronse 3 volumes: *Gallego 1* (1971), *Gallego 2* (1972) e *Gallego 3* (1974).

[6] De feito, en 1977 publícanse as *Bases prá unificación das normas lingüísticas do galego*, resultado dunha serie de seminarios impulsados polo ILG.

cada que permita o seu uso nos diferentes estamentos da sociedade galega[7]. Apróbanse así en 1982 as *Normas ortográficas e morfolóxicas do idioma galego* e, posteriormente, en 1995 e en 2003 lévanse a cabo algúns cambios. En consecuencia, por un lado primeiramente a ausencia de normativa e logo as modificacións posteriores, e polo outro a situación de dúas linguas en contacto —galego e castelán— conducen a unha enorme variación gráfica e morfolóxica que caracteriza en xeral os textos do CORGA, fundamentalmente os procedentes dos 70 e 80, dado que, como xa sinalamos, respéctase a solución escollida polo autor, sen corrección de grallas nin estandarización de ningún tipo[8].

O CORGA enriqueceuse coa etiquetaxe morfosintáctica automática dos seus textos, o que permite dar un salto cualitativo na recuperación de información e supera-las restricións que impoñen as consultas baseadas en formas ortográficas. Así, é posible realizar procuras tanto por palabras ortográficas como por datos lingüísticos (elemento gramatical, clase de palabra, etiqueta, lema, hiperlema e/ou trazos morfolóxicos) segundo diversos criterios combinables entre si: período temporal, área temática, medio, tipo de documento, parte do documento etc. Para facer posible esta recuperación, no CRPIH desenvolveuse o etiquetador/lematizador XIADA, unha ferramenta para PLN coa que se etiqueta automaticamente todo o corpus (40 178 271 palabras ortográficas / 48 184 012 elementos gramaticais). Dada a enorme variación que presentan os documentos que conforman o corpus, o reto é tentar recoñecer e analizar non só aquelas unidades que coinciden coa normativa, senón tamén as variantes gráficas e morfolóxicas non estándares ou solucións orixinadas polo contacto de linguas que son usuais nos textos.

O contido deste artigo está estruturado como segue. No apartado 1 trázanse as características xerais do CORGA. No apartado 2 descríbese o deseño, composición e elaboración; o apartado 2.1 céntrase no deseño, o 2.2 amosa os datos sobre a composición e o 2.3 describe o proceso de construción e incorporación dos documentos ó corpus. No apartado 3 explícanse as características do sistema de codificación que se emprega. No 3.1 dis-

[7] Lembremos que en 1981 se aproba o Estatuto de Autonomía no que se recoñece a lingua galega como oficial na comunidade.

[8] Cómpre matizar lixeiramente este aserto, pois nos textos de procedencia oral si levamos a cabo certa estandarización, por exemplo á hora de representa-las asimilacións, mais hai que ter en conta que non partimos dun texto escrito. Somos nós quen vertemos á escrita o rexistro sonoro, ou sexa, transcribimos, o que nos obriga a tomar decisións sobre como representar certos fenómenos lingüísticos. Pola contra, nos textos de procedencia escrita esas decisións xa as tomaron os autores e nós respectámolas.

cútense as marcas de codificación utilizadas no corpus e no 3.2 descríbese a estruturación dos distintos tipos de documentos. O apartado 4 está adicado á anotación morfolóxica e á explicación do funcionamento do sistema XIADA, do que sobresaen, tanto pola gran cantidade que presenta o galego como pola complexidade do seu recoñecemento, as contraccións e mailas formas verbais con pronomes enclíticos. Por último, no apartado 5 descríbese polo miúdo o sistema de consultas que se deseñou para a recuperación de información.

2. Deseño e composición

2.1. O deseño do corpus e a súa clasificación

Entendemos que para que un corpus sexa representativo do uso lingüístico dunha lingua dada nun período concreto debe existir un deseño previo sobre os documentos que vai acoller. Neste sentido as premisas que guían o deseño do rexistro escrito do CORGA son: i) outorgar maior representatividade ós períodos máis recentes e menos ós máis antigos[9]; ii) equilibrar na medida do posible o tamaño do corpus entre a ficción e a non ficción, inda que polas características da propia lingua sexa máis doado atopar textos ficticios ca ensaísticos ou xornalísticos; iii) representar tódalas áreas temáticas nos distintos períodos cronolóxicos nos que se estrutura o corpus, e iv) dar cabida a un abano amplo de autores.

Así, os parámetros que se teñen en conta para a selección dos textos son a *data,* o *tipo de texto* e a *área temática* coa que se clasifica o documento. Respecto da *data*, o deseño realízase tomando en consideración períodos de 5 anos —os lustros completados ata o de agora son 1975-1979, 1980-1984, 1985-1989, 1990-1994, 1995-1999, 2000-2004, 2005-2009, 2010-2014 e está inconcluso o lustro 2015-2019—, mentres que segundo o *tipo de texto* selecciónanse xornais, revistas, ensaios, novelas, relatos curtos, obras de teatro e, na última década, tamén guións e blogs. Por último, a escolla dos documentos de non ficción —xornais, revistas, ensaios e blogs— persegue representar tódalas áreas e subáreas temáticas recollidas na figura 1:

[9] En parte débese á dificultade de atopar textos escritos para algunha área temática ou tipo de documento concretos e en parte entendemos que interesa máis reflecti-la lingua dende o momento da súa oficialidade.

	Áreas temáticas				
	Economía e política	**Cultura e artes**	**Ciencias sociais**	**Ciencias e tecnoloxía**	**Outros**
Subáreas temáticas	Política	Audiovisuais e espectáculo	Lingua	Sanidade	Deportes
	Desenvolvemento e infraestruturas	Medios de comunicación	Literatura	Bioloxía, botánica, ecoloxía, zooloxía e paleontoloxía	Turismo
	Emprego, traballo, industria	Artes gráficas e plásticas	Relixión	Tecnoloxía e industria	Afeccións e asuntos domésticos
	Sector servizos	Patrimonio, arquitectura, arquivos	Historia e xeografía	Medio, astronomía e xeoloxía	Actualidade, sucesos, homenaxes, inauguracións
	Explotación primaria		Civilización, etnoloxía, arqueoloxía e antropoloxía	Matemáticas e estatística	Biografía
	Economía, facenda, bolsa		Pensamento, ética e filosofía	Química, bioquímica e farmacia	Nota prologal
	Ordenación sanitaria		Socioloxía e psicoloxía		
	Xustiza, lexislación, dereito		Erotismo e sexoloxía		
	Asuntos sociais		Astroloxía e ocultismo		
	Ordenación académica				

FIGURA 1. Valores para o parámetro de clasificación *área temática*

Amais dos criterios utilizados para a selección dos documentos que abrangue o CORGA —*data*, *tipo de texto* e *área temática* para os textos xornalísticos e ensaísticos—, sérvennos para clasificar tipoloxicamente tódolos textos que se introducen no corpus as variables *orixe, bloque, xénero* e *subtipo*, as cales sucesivamente dan conta da procedencia do documento, da súa pertenza ó bloque da ficción ou non ficción, do xénero no que se inscribe o texto e do subtipo concreto de documento ante o que estamos. Na parte oral adxudícanse só as variables de clasificación *orixe* e *subtipo*, fronte á parte escrita na que se aplican todas. A modo de sinopse, recollemos na figura 2 os valores posibles para cada unha das variables catalogadoras da tipoloxía textual e as dependencias que se establecen entre elas:

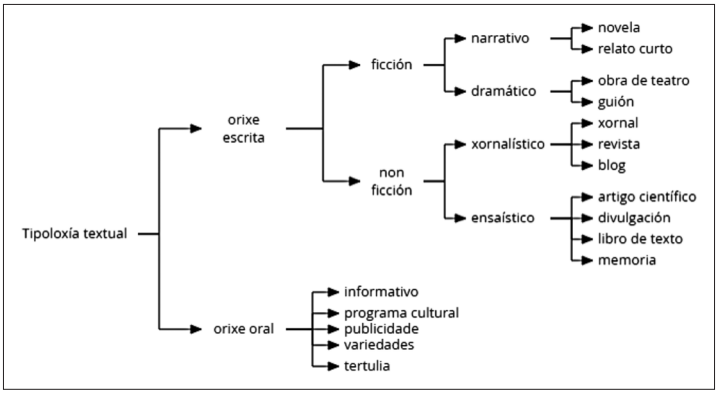


FIGURA 2. Valores e dependencias posibles para a tipoloxía textual

2.2. Datos sobre a composición do corpus

A versión 3.1 do CORGA consta de 40 178 271 palabras que se distribúen do seguinte xeito:

— 641 xornais, que representan 12 181 778 palabras distribuídas en 41 122 noticias. Inclúe exemplares completos d'*A Nosa Terra, A Peneira, De Luns a Venres, Diario Oficial de Galicia, Galicia Hoxe, Novas do Eixo Atlántico, O Correo Galego, O Xornal de Galicia* e *Sermos Galiza*.

— 293 revistas, que supoñen 5 514 372 palabras repartidas en 8502 noticias. Engloba números d'*A Voz de Vilalba, Cerna, Código Cero, Consumer Eroski, Disquecool, Díxitos, Entregas de Comunicación Cultural, Feiraco, Galicia Internacional, Gciencia, Luzes, Man Común, Petroglifo, SCQ Basket Magazine, Sprint motor, Teatro do Noroeste, Teima, Tempos Novos* e *Revista Galega de Economía*.

— 569 libros distribuídos en 2809 documentos que totalizan 21 447 180 palabras repartidas do seguinte modo[10]:

  — 189 novelas, que suman 9 671 594 palabras.
  — 161 ensaios, que supoñen 7 832 692 palabras distribuídas como segue: 7 libros de texto cun total de 511 898 palabras; 10 memorias que globalmente computan 474 923 palabras; 52 artigos científicos cun global de 321 505 palabras e 617 documentos de divulgación que suman 6 524 366 formas.
  — 129 coleccións de relatos, que supoñen 2 829 955 formas ortográficas.
  — 91 obras de teatro, que conforman un global de 1 312 261 palabras.

— 7 guións televisivos (*A vida por diante, As leis de Celavella, Matalobos, Padre Casares, Pazo de familia, Serramoura* e *Terra de Miranda)* e 1 incluído nunha colección de relatos (*A noite dos cans doentes)* que suman 78 082 palabras.

— 157 entradas constituídas por 65 864 palabras pertencentes ós blogs *Natureza Dixital*, *Se o vello Sinbad tivese fillos* e *Está en galego!*

— 25 horas de transcricións dos programas radiofónicos *Camiño de Volta, Crónica das dúas, Diario Cultural, Informativo, O miradoiro. Paco Feixó, Pensando en ti* e *Xamón Xamón*. En total supoñen 263 407 palabras.

[10] Ante a imposibilidade de relaciona-los títulos, remitimos á listaxe recollida no enderezo http://corpus.cirp.gal/corga/datos?subcorpus=etiquetado+automaticamente.

Dende a xénese do CORGA, o rexistro oral dispuña dun espazo propio no deseño do corpus[11], mais a enorme desproporción entre os esforzos adicados e o reducido volume de texto transcrito, amén das dificultades computacionais existentes naquela altura para o seu tratamento, provocaron que nos centrásemos primeiramente no rexistro escrito e procrastinásemo-lo oral. E só cando estivesen encarreirados, definidos e testados os procedementos de introdución dos textos escritos, cando fósemos capaces de tratar computacionalmente a variedade escrita cun mínimo de rigorosidade, só entón retomariámo-las transcricións. Esta é a razón de que non estean representados tódolos períodos temporais para o rexistro oral, de que os criterios de selección respondan á variedade escrita e de que, en definitiva, se evidencie tal desequilibrio entre a procedencia escrita e maila oral.

Na actualidade estanse transcribindo os rexistros sonoros da década presente e paulatinamente iranse completando os datos para as demais décadas.

2.3. A elaboración do corpus

Nunha primeira fase que abrangueu aproximadamente ata 2011 os documentos dixitalizábanse a través de escáner e logo aplicábaselles un recoñecedor óptico de caracteres (OCR). Para a dixitalización dos textos empregamos dende o ano 1995 ata o 2008 o escáner Fujitsu M3096GX e xa a partir do 2008 o modelo HP Scanjet N9120. Respecto do software de recoñecemento, utilizamos distintas versións de OmniPage, dende a OmniPage Pro dispoñible en 1995 ata a actual OmniPage Ultimate 19.1.

Senón as novas d'O Correo Galego, que grazas a un convenio existente chegaban a nós xa en formato electrónico, ata que empezaron a existir na rede versións electrónicas de xornais e revistas, o paso do documento en papel ó arquivo electrónico era un labor arduo e proclive á introdución de erros no proceso de OCR, dado que nos primeiros anos non existía soporte para a lingua galega. Non obstante, dende o ano 2011, grazas á xenerosidade das editoriais e ó auxe de Internet como lugar de depósito consultable de numerosas publicacións, o soporte papel desaparece do noso procedemento. Con todo, permanece o OCR de OmniPage para a delimitación de noticias en revistas cando partimos dun arquivo en pdf, ou para aqueloutros documentos, tamén en pdf, que presentan problemas para a súa exportación a un arquivo de texto copiando e pegando o contido do documento.

[11] Proba disto é que o primeiro material sonoro que se incorpora ó corpus corresponde a programas radiofónicos da década dos 90, data na que bota a anda-lo proxecto.

Os formatos de partida son maioritariamente pdf e html, e rara vez doc, odt ou txt. O formato de chegada é nun comezo o de só texto, para o que empregámo-lo editor de texto TextPad. Na versión txt estrutúrase e codifícase o documento orixinal, aplicando as distincións recollidas no apartado *3. A codificación e estruturación,* conforme un complexo protocolo que ten en conta as particularidades dos distintos tipos de documentos. Agora ben, esta fase presenta unha peculiaridade e é que, en función do tipo de documento do que se trate, a través da organización en carpetas, da existencia de arquivos paralelos, un para o texto e outro para os metadatos, e por último a través da propia organización do texto no arquivo, por medio duns *scripts*, os arquivos convértense ó formato xml, no que a estrutura trazada mediante a disposición dos arquivos na carpeta e a propia disposición interna dos arquivos en txt transfórmase agora en texto delimitado por etiquetas xml (Barcala 2010).

Moi resumidamente e exemplificando cun número do xornal *Sermos Galiza*, a partir do formato pdf delimítanse as noticias constituíntes a través do software de OmniPage de forma que cada noticia conforma un arquivo (noticia_1, noticia_2 etc.). Previamente, nunha carpeta co nome do xornal, o ano, o mes e mailo día créase unha carpeta para cada unha das seccións nas que se clasifican as noticias (Opinión, Deportes, Economía, Internacional, Galicia...), e cunha numeración consecutiva vaise introducindo cada noticia delimitada na sección que corresponda. Unha vez delimitadas todas, procédese á revisión do texto, á introdución de etiquetas se for preciso, e á súa fragmentación, se é o caso, en *titular, corpo* e *pé de foto*. Logo, cando o texto da noticia está revisado, codificado e fragmentado, en paralelo, créase un arquivo coa mesma numeración pero co identificador *cabeceira* (cabeceira_1, cabeceira_2 etc.) no que se inclúen os metadatos específicos da noticia (identificador, autor, ata tres áreas temáticas e un campo comentarios se houber algo que consignar). Rematada a revisión de tódalas noticias coa construción de cadansúa cabeceira, constrúese a cabeceira para o xornal completo, na que se inclúe o identificador, o nome, a clasificación textual, o soporte, a editorial, o lugar de publicación, o depósito legal etc.

Posteriormente, esta estruturación en múltiples niveis de arquivos e carpetas trasládase mediante a aplicación duns *scripts* a un arquivo en formato xml que contén tódalas noticias que conforman o xornal, organizadas sucesivamente en *cabeceira* e *corpo*, e este último estruturado ademais nas súas partes constitutivas. O arquivo en xml ofrece outra particularidade: os parágrafos foron segmentados automaticamente en secuencias menores separadas por un signo forte de puntuación (punto, punto e coma, dous puntos...).

Logo, sobre esta versión en xml, empregando o editor XMLmind XML Editor, ó que se lle engadiu unha capa de personalización da visualización para face-la manipulación dos documentos máis doada, unha lingüista revisa novamente o texto completo comprobando que se corresponde co orixinal, supervisa a estruturación e maila segmentación, introduce as referencias ás notas e as notas mesmas se o documento as posúe, as etiquetas de táboa e fórmula, subíndice ou superíndice e especifica os valores das etiquetas introducidas na fase previa alí onde corresponda. Unha serie de ficheiros DTD (definición de tipo de documento) describen a estrutura dos diversos tipos de documentos que contén o CORGA (*blog, colección, comúns, guión, libro, oral, revista, teatro* e *xornal*), de xeito que a revisión do arquivo xml conta sempre cunha DTD que garante que a estrutura do arquivo é congruente coa estrutura deseñada e, así mesmo, sérvelle de guía ó revisor, por canto restrinxe a parte do texto na que é posible introducir etiquetas, os valores que estas poden levar asociado ou a orde na estruturación.

No que respecta á procedencia oral, a transcrición dos audios realízase co programa ELAN que permite aliña-lo texto co son. Para as transcricións dos programas de radio da década dos 90 o procedemento foi totalmente manual, pero na actualidade, tras asinar un convenio de colaboración co Grupo de Tecnoloxías Multimedia da Universidade de Vigo[12], empregámo-la súa plataforma de subtitulado para obter unha versión preliminar da transcrición, a cal proporciona o texto transcrito aliñado co son nun único segmento no formato eaf, editable xa que logo co ELAN. Posteriormente esa liña divídese en tantas liñas como falantes se rexistren, corríxense os erros de transcrición, delimítanse en secuencias as quendas de palabra, introdúcense as pausas e, en xeral, codifícase o texto de acordo co protocolo de traballo establecido.

A transcrición é ortográfica, polo que non se reproducen casos de gheada, seseo, rotacismo, paragoxe etc. Entre as asimilacións que se realizan represéntanse exclusivamente as que recolle a norma oficial (*pero* non **pro*; *contra o* non **contró*; *para o* e non **pró* ou **pó*; *Dios os dá* e non **Dio-los dá*; *non o fixen* pero non **nono fixen*), mais si reflectimos no caso de existir máis dunha alternativa de representación a que realiza o falante (*de algún dos membros vs. dalgún atentado*; *te-la pel branca vs. ter a sensación*). Non se estandariza a morfoloxía, de xeito que se transcribe, entre outras, a acentuación proparoxítona na P4 e P5 (*estábamos, rematábamos*), plurais en *-s* no sitio do estándar *-ns* (*millós*, *acciós*, *ocasiós*), a apócope dos substantivos en *-ade* (*verdá, necesidá*), variantes no vocalismo (*millor, pior, muito, mismo*),

[12] <http://gtm.uvigo.es/content/speech-technology>.

castelanismos (*ahora, hasta, hastra, Dios, jueves*) etc. Por outra banda, o emprego da maiúscula limítase ás iniciais dos nomes propios e mais ás siglas e non se usan signos de puntuación, substituíndoos a etiqueta *pausa,* fóra o de peche de admiración ou interrogación.

Para evitar erros na medida do posible, nos arquivos de procedencia escrita o creador do documento txt diverxe do revisor da versión xml, e nos de procedencia oral as transcricións pasan por tres pares de ollos diferentes.

3. A codificación e estruturación

Deseguido imos ver que tipo de actuacións se practican sobre os documentos que se incorporan ó corpus, centrándonos na codificación do texto propiamente dito. Comezamos coa análise das marcas que dan conta de fenómenos lingüisticamente relevantes e seguimos no apartado 3.2 cos factores que condicionan a propia existencia do documento e cuxa disposición interna se traslada a partes estruturais en función do tipo concreto de texto do que se trate.

3.1. Marcas de codificación

Os documentos que se incorporan ó CORGA non manteñen a paxinación nin os formatos presentes nos orixinais, de xeito que as indicacións de páxina, sangrías, cursivas, grosas ou versaletas desaparecen no proceso de integración sen deixar pegada. Mais, en contrapartida, aplícaselles ós textos unha marcaxe que informa sobre algunhas características lingüisticamente relevantes a través de etiquetas xml de apertura e de peche que envolven o fragmento textual afectado, coa única condición de que a unidade mínima sobre a que se aplica é a palabra, dado que unha etiqueta non pode rompe-la unidade léxica ou, en caso contrario, dificultaríase o seu procesamento automático.

As etiquetas empregadas na codificación dos textos transfórmanse no sistema de recuperación de información en palabras sobre fondo amarelo ou branco, de xeito que se facilita a información cun deseño amigable —a cor codifica unha etiqueta cuxa identificación individual emerxe ó situa-lo cursor enriba da palabra destacada— e evítase o inconveniente de te-las etiquetas no medio do texto, facendo que a lectura dos resultados sexa fluída.

En (1) ofrécese como mostra a codificación dun segmento en texto plano e na figura 3, de esquerda a dereita e de arriba cara abaixo, respectivamente, a visualización do fragmento no pdf do que partimos —o destacado é noso para sinalar que é xusto nese segmento onde aplica a etiqueta—, a codificación no editor de texto que empregamos para a inclusión dos documentos —o

XMLmind XML Editor Personal Edition— e finalmente como se fai visible no resultado da aplicación de consulta:

(1) <oración>É do xénero tonto pensar que a autodeterminación de Euskalherría ha de votala a España cañí.</oración>
<oración>A este paso, como letra do himno español adoptarán aquel estribillo que moito cantaba o amigo Temes nas troulas universitarias:</oración>
<oración distinto=»outra_lingua»>¡Qué tontería, pensaba que tú eras mía!»</oración>

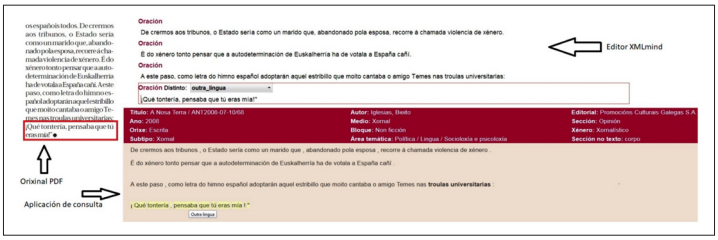


FIGURA 3. Mostra da visualización da etiqueta *outra_lingua* no editor XML e na aplicación de consulta

As etiquetas que se aplican na codificación dos textos, e que imos relacionar deseguido co nome da marca que é visible para o usuario, deseñáronse cun triplo obxectivo:

(i) Impedir que texto que non forma parte do acervo léxico galego actual produza ruído no cómputo de formas e mais na recuperación da información, ben porque son formas doutras linguas[13], ben porque pertencen a un período anterior do galego, ben porque se trata de erros na dicción ou de palabras cortadas. Con este propósito incorpóranse as seguintes etiquetas:

— Descoñecido	Fragmento non identificable con ningunha lingua.
— Non normativo	Texto que segue unha normativa de mínimos.
— Outra lingua	Fragmento nunha lingua distinta do galego.

[13] Márcanse os parágrafos e secuencias integramente escritos nunha lingua distinta do galego, sen especificar de que lingua se trata. Agora ben, as palabras soltas non se marcan (*mass media, spoiler, modus operandi, cool…)*, a non ser que apareza ó lado a tradución ou o texto especifique explicitamente que se trata dunha unidade pertencente a outra lingua. Para unha descrición máis detallada, véxase http://corpus.cirp.gal/corga/etiquetas_codificacion

— Outro período	Texto en galego anterior a 1955.
— Palabra cortada	Fragmento dunha palabra.
— Sic	Erro de dicción (aplícase só nas transcricións).

As formas caracterizadas con estas marcas non poden ser obxecto de procura, non se teñen en conta para a elaboración das diversas listaxes dispoñibles na pestana *Descargas* nin se anotan morfoloxicamente. Só son visibles como contexto doutra busca.

(ii) Indicar que no lugar onde se atopa hai texto que está disposto en filas e columnas ou consta de símbolos matemáticos, físicos, químicos etc. que non son representables tendo en conta as posibilidades de estruturación que definimos ou, no caso das transcricións, que non resulta intelixible:

— Fórmula	Fórmula non recollida no corpus. Pode substituír tamén algún carácter grego ou cirílico.
— Táboa	Táboa non recollida no corpus.
— Inintelixible	Fragmento incomprensible e, polo tanto, non transcrito.

(iii) Proporcionar información sobre a representación gráfica do segmento marcado. Das etiquetas de codificación recollidas neste subapartado unhas afectan unicamente os textos de procedencia escrita —*subíndice, superíndice, texto en táboa, texto riscado* e *texto unido*—; outras utilízanse só nas transcricións —*rindo, risa, ruído bucal, texto solapado, transcrición dubidosa* e *transcrición non documentada*— e, por último, outras úsanse tanto no rexistro oral coma no escrito —*alongamento* e *texto en verso*—.

— Alongamento	No rexistro oral, aumento da cantidade dalgún son; no escrito, reduplicación de letras ou sílabas
— Rindo	Texto emitido entre risas.
— Risa	Substitúe a onomatopea da risa cando esta non afecta a alocución.
— Ruído bucal	Substitúe os distintos ruídos bucais que pode emitir un falante: chasquidos, aclaración da voz, rouquén etc.
— Subíndice	Segmento en subíndice.
— Superíndice	Segmento en superíndice.
— Texto en táboa	Texto presente nunha táboa.
— Texto en verso	Texto pertencente a un poema ou unha canción.

— Texto riscado	Palabra ou fragmento riscado.
— Texto solapado	Texto emitido simultaneamente con outro.
— Texto unido	Texto soletreado, palabras unidas con trazos ou fragmento textual corrido.
— Transcrición dubidosa	Transcrición comprendida, mais non totalmente segura.
— Transcrición non documentada	Transcrición cuxa representación gráfica non foi posible documentar[14].

En tódolos casos limitámonos a situar dentro da etiqueta xml o texto que se ve afectado pola característica que indica cada unha das etiquetas, ou ben indicamos que onde está a etiqueta existía unha táboa ou unha fórmula, mais non se corrixe, normativiza nin manipula a grafía dos documentos. Agora ben, hai dous casos nos que si alterámo-la grafía, terxiversando polo tanto o texto orixinal. Trátase de *alongamento* e *texto unido*. A primeira é unha etiqueta que se incorpora nos textos escritos a raíz da inclusión de transcricións ortográficas no corpus, aplicándoa dende a versión 3.0 nos textos de nova inserción e paulatinamente nos xa incorporados, mentres que a segunda emerxe coa incorporación dos blogs e xorde para dar cabida á nova alternativa de formato que ofrecen algúns procesadores de texto: o riscado[15].

O enriquecemento do corpus coa etiquetaxe automática de tódolos seus documentos é o causante desta manipulación, pois nos textos escritos os autores recorren á reduplicación de vogais e consoantes (*veeeennnnn*), á repetición de sílabas (*fafafafacer* ou *mi-milagre*), ós soletreos (*ab-so-lu-ta-men-te*) ou á aglutinación de palabras (*son-un-ho-me-feliz*) para indicar intensidade na alocución dun personaxe, caracteriza-la tartamudez, un estado de embriaguez etc. Isto orixina que esas unidades, recoñecibles para un humano como elementos pertencentes ó paradigma de *facer, milagre, absolutamente* etc., se converten en unidades descoñecidas para unha máquina, sen que esta poida establecer unha relación entre a forma sen alongamento e con el. Pois ben, para facilita-lo recoñecemento automático,

[14] Esta etiqueta emprégase con palabras que non sabemos se transcribimos ben, mais que non son inintelixibles, por exemplo *Koderanka* no fragmento «a partir das dez e media da noite <pausa/> o ballet moldavo <transcricion_non_documentada>Koderanka</transcricion_non_documentada> <pausa/> i a Orquesta Sinfónica Nacional de Moldavia <pausa/> representarán na Praza da Igrexa de Lalín <pausa/> o espectáculo Carmelu» [CORGA: Informativo 14:00. 24/07/1997].

[15] A diferenza da negra ou a itálica, o riscado non só é un formato, pois debaixo del existe texto; ou sexa, persiste o rastro de que houbo unha proposta anterior que, semántica e sintacticamente é equiparable á actual, configurando ambas unha especie de relación paradigmática.

no texto que se incorpora ó corpus modifícase a forma para a súa grafía convencional: suprímense os caracteres ou sílabas reduplicados e elimínanse os trazos indicativos do soletreo, pero permanece a súa pegada ó traslada-la indicación de *alongamento* e *texto unido* á etiqueta.

Existen aínda outras dúas etiquetas que son visibles só dende o contexto, están destacadas sobre fondo branco e, o máis significativo, permiten centra-las buscas no texto que clasifican. Trátase de *acoutación* e *interlocutor,* ambas relacionadas con tipos de documentos específicos, o que nos conecta directamente co apartado seguinte.

3.2. A estruturación dos documentos

Os documentos que se incorporan ó CORGA codifícanse segundo o estándar XML (*eXtensible Markup Language*) co fin de incrementa-las posibilidades de recuperación de información e, ademais, garanti-la permanencia no tempo. Esta codificación implica un deseño e estruturación que dá conta da disposición interna característica de cada un dos grandes tipos de textos (*xornal, teatro, ensaio* etc.). Así, exemplificando cun texto xornalístico, esta estruturación permite considerar un xornal un único documento que está organizado en noticias, distribuídas en seccións, as cales, á súa vez, conteñen obrigatoriamente un *corpo* e opcionalmente un *titular, resumo* e/ou *pé de foto*. A maiores, cada un destes elementos está constituído por parágrafos (texto comprendido entre dous puntos e á parte), e estes son segmentados en enunciados menores (secuencia textual separada do resto do texto por un signo forte de puntuación). Naturalmente, o xornal posúe ademais unha cabeceira complexa na que se recollen os datos bibliográficos e unha cabeceira específica por noticia onde se inclúen os datos relativos ó seu autor, a sección do xornal na que se localiza e ata tres áreas temáticas que a clasifican en canto ó seu contido.

Esta disposición detallada habilita a posibilidade de, no sistema de recuperación de información, unha vez indexados os documentos, realizar consultas sobre a totalidade do documento (noticia, seguindo co exemplo) ou sobre unha unidade estrutural concreta (para o tipo de documento *xornal*: *titular, resumo, pé de foto* ou *corpo*).

Pola súa banda, no medio *libro,* o texto estrutúrase en *partes* e *capítulos*, ou nas obras de teatro e guións en *actos*, *cadros* ou *escenas*, en ámbolos dous casos se os houber, ou senón en unidades que denominamos *división* —necesariamente todo documento posúe como mínimo unha—, o que nos permite a ulterior separación en *encabezamento* e *corpo* de cada unha das partes existentes, o primeiro potestativo e o segundo obrigatorio.

Inflúen tamén na estruturación do documento a autoría e a nuclearidade. Con respecto á primeira, a existencia de autores diferentes dá lugar a documentos diferentes. Así, unha colección de relatos ou unha obra de teatro adoita estar prologada por un autor distinto do autor dos relatos ou peza teatral, inda que forme parte fisicamente do mesmo libro, o que os converte en documentos distintos para o sistema, e en consecuencia leva asociada unha cabeceira na que se recollen os metadatos e se dá conta do aniñamento. O mesmo sucede con calquera obra que reúna distintos textos de distintos autores, xa for obras de teatro, xa for relatos, xa for artigos científicos ou xornalísticos, de aí o alto número de «documentos» que se observan nos datos das frecuencias do CORGA.

Con respecto á nuclearidade do documento e tendo en conta a disposición orixinal, tentamos discriminar na estruturación entre os elementos constitutivos periféricos —*prólogo, dedicatoria, cita, nota* e *apéndice*— do elemento constitutivo nuclear —o *corpo*—. As unidades estruturais *prólogo* e *apéndice* engloban as distintas denominacións que acostuman aplicárselles nos textos: *presentación, limiar, introdución, preámbulo* ou *prefacio* no caso do *prólogo* e *epílogo, agradecementos, táboa de feitos históricos, glosario* ou *vocabulario* no caso de *apéndice*.

Por último, a medio camiño entre as marcas de codificación e as unidades estruturais propias dun tipo de texto concreto sitúanse as indicacións de *acoutación* e *interlocutor*. A primeira, *acoutación,* delimita o discurso contido nos diversos tipos de didascalias das obras teatrais e guións, mentres que a segunda, *interlocutor,* asigna o texto a un personaxe concreto dunha obra de teatro ou guión, un interveniente nunha entrevista, coloquio, mesa redonda etc., o locutor dun informativo, un participante nunha tertulia ou calquera tipo de falante dunha transcrición. A finalidade primixenia é eliminar do texto os nomes dos personaxes para que non terxiversen os datos léxicos do galego interferindo nas frecuencias[16]. Así, o nome do personaxe recollido explicitamente no texto orixinal, ou discriminado polo formato, seguido en xeral de dous puntos ou raia, é eliminado e convertido en etiqueta no documento que se incorpora ó corpus. Mais, posto que o fin último que perseguimos é posibilita-la obtención da maior cantidade de información posible dos documentos que contén o CORGA, *interlocutor* e *acoutación,* para os efectos de recuperación de información, compórtanse como partes

[16] A modo de exemplo, pénsese en como a aparición de, supoñamos, 300 casos de Avelino, 300 de Pousa e outros 300 de Antelo na listaxe de formas completas, cuxa presenza se debería maioritariamente á introdución no CORGA do ensaio *Conversas con Avelino Pousa Antelo,* levaría a interpretar que Avelino, Pousa e Antelo son antroponímicos frecuentes en galego.

estruturais, de xeito que a consulta pode centrarse só nas *acoutacións* ou focalizala mediante *interlocutor* no rexistro oral, tanto procedente do soporte escrito coma do soporte auditivo.

4. A ETIQUETAXE

O CORGA está etiquetado automaticamente na súa totalidade co Etiquetador do galego actual XIADA, ferramenta desenvolta conxuntamente entre o Centro Ramón Piñeiro para a investigación en humanidades e o grupo COLE das universidades da Coruña e Vigo.[17] Trátase dun etiquetador estatístico baseado nos modelos de Markov de grao 2 que se alicerza nuns recursos lingüísticos, descritos polo miúdo en Domínguez (2013), dos que deseguido imos ve-las características máis salientables.

O etiquetario empregado na anotación morfosintáctica do corpus definiuse tendo en conta as recomendacións de EAGLES, o camiño percorrido xa na etiquetaxe de corpus en linguas coma o portugués, o catalán ou o castelán e, principalmente, fundamentándonos nos estudos gramaticais dispoñibles para o galego. O etiquetador traballa con 382 etiquetas diferentes, resultado de combina-las 18 clases de palabras recoñecidas cos valores das categorías gramaticais pertinentes en cada unha delas: substantivo (16), adxectivo (21), verbo (75), preposición (1), conxunción (2), adverbio (4), artigo (8), demostrativo (9), relativo (11), posesivo (56), indefinido (18), numeral (18), pronome (86), interrogativo-exclamativo (14), locución (4), interxección (1), sinal de puntuación (20) e 18 para a categoría periférica (discrimínase nesta clase a través do subtipo entre abreviatura, sigla, fórmula, símbolo e outros tipos). Dende a versión 3.1, debido á presenza no corpus de numerosas formas con grafías innovadoras para a expresión do xénero nunha linguaxe que pretende ser non sexista (Domínguez e Barcala 2018; Caíña, Domínguez e López 2019), o etiquetador traballa cun total de 453 etiquetas, 71 máis cás que acabamos de relacionar, resultado de implementar no sistema XIADA o valor *xenérico* para a categoría gramatical xénero naquelas clases de palabra nas que é pertinente: substantivo (2), adxectivo (4), verbo (2), artigo (4), demostrativo (4), relativo (4), posesivo (24), indefinido (4), numeral (4), pronome (10), interrogativo-exclamativo (4) e categoría periférica (2).

O sistema de etiquetaxe de XIADA presenta unha estrutura xerárquica na que no primeiro nivel se identifica a clase de palabra e nun segundo esténdense os atributos gramaticais pertinentes para cada clase. A distribución

[17] A última versión deste etiquetador está dispoñible na páxina http://corpus.cirp.gal/xiada tanto para a súa descarga, xunto cos recursos que emprega, como para o modo demostración.

dos atributos é simétrica, de tal xeito que tódolos elementos pertencentes a unha clase dada posúen o mesmo número de atributos e os seus valores ocupan a mesma posición dentro da cadea resultante. Así mesmo, a etiquetaxe, que en xeral segue os manuais da lingüística galega, presenta unha peculiaridade que a distancia da práctica habitual e é que se asigna o valor máis específico que permite o contexto da unidade en cuestión dentro da secuencia (delimitada por signos fortes de puntuación cuxa segmentación se realiza automaticamente e logo é supervisada polo equipo lingüístico, como vimos no apartado 2.3). Así, por exemplo, as etiquetas dispoñibles no sistema para o adxectivo *amable* son:

— A0as: adxectivo, grao non aplica, masculino/feminino, singular
— A0fs: adxectivo, grao non aplica, feminino, singular
— A0ms: adxectivo, grao non aplica, masculino, singular

Pois ben, na etiquetaxe do corpus, *amable* é tratada, entre outros moitos casos, como forma feminina en *Agora desexaría sabe-la razón da súa amable visita*, masculina en *Non se trataba dun olor amable, delicado, grato, insinuante, pracenteiro, suave…* e ambigua en canto ó xénero en *Non me costaba nada ser amable e decirlle que estaba ben, que a botaba de menos a pesar de que nunca me escribía*.

Unha parte do corpus etiquetouse automaticamente e logo supervisouse dun xeito manual a súa desambiguación comprobando que cada unidade gramatical recibise a etiqueta morfosintáctica e o lema que lle correspondía segundo o contexto no que se localizaba; nun primeiro momento foron novas do xénero xornalístico e logo parágrafos extraídos aleatoriamente de relatos curtos. Este subcorpus, do que se puxeron á disposición pública varias versións co nome de *Corpus de Referencia do Galego Actual etiquetado (CORGAetq)* e que hoxe está dispoñible baixo a opción *etiquetado manualmente* na aplicación de consulta do CORGA, consta de algo máis de 600.000 palabras ortográficas e foi o que se empregou para adestra-lo etiquetador.

O corpus está etiquetado integramente, ou sexa, dos 48 millóns de elementos gramaticais que posúe o CORGA na súa versión 3.1, tódalas unidades léxicas que figuran nos textos, sexan estas vocábulos pertencentes ó léxico común, sexan entidades, datas, cifras, siglas, abreviacións, símbolos, url, enderezos de correo electrónico, identificadores de enumeracións etc., reciben unha caracterización morfosintáctica. A día de hoxe descoñecémo-la taxa de acerto que posúe o etiquetador, pois o dato coñecido dun 96% de precisión

que nos consta (Domínguez *et al.* 2009) refírese unicamente á etiquetaxe realizada sobre textos xornalísticos. É pois unha tarefa pendente inquirila.

O proceso de etiquetaxe resúmese na figura 4, onde a caixiña central destacada en vermello constitúe o punto de destino dos demais recursos que entran en xogo e que imos sintetizar brevemente. O preprocesador delimita as unidades léxicas presentes nun texto e asígnalles tódalas análises posibles tendo en conta os datos que figuran no lexicón para as formas simples, e os que constan nas táboas externas para as contraccións e os pronomes enclíticos. Así mesmo, tendo en conta os datos que lle achega o corpus de adestramento e as regras, o etiquetador escolle a análise que lle parece a correcta segundo o contexto no que se sitúa o elemento gramatical.

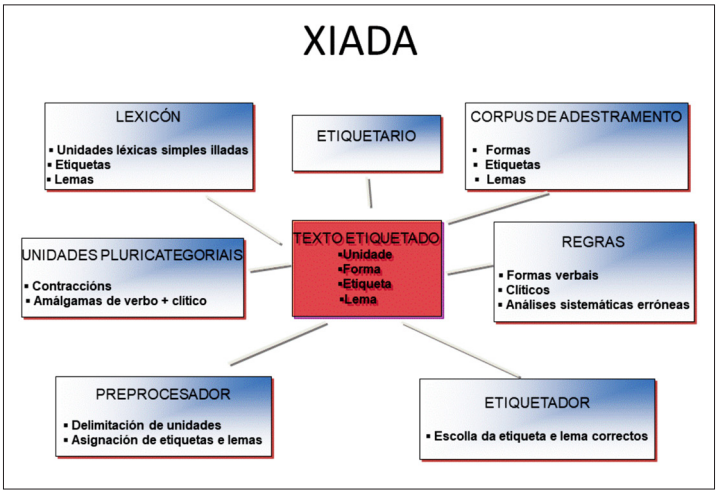


FIGURA 4. O sistema XIADA. Gráfico elaborado a partir de Domínguez (2013)

No lexicón intégrase a información morfosintáctica que lle corresponde a cada unidade gramatical simple illada no texto, sen incluí-las contraccións ou conglomerados de forma verbal e enclíticos, cuxo recoñecemento se produce por outro lado. Así pois, no lexicón non se recollen, por exemplo, os alomorfos *lo* e *no* do pronome de terceira acusativo, xa que estes se amalgaman sempre a outras unidades, pero si a primeira forma *o*. É no lexicón xa que logo onde se indica que *amable* é un adxectivo no que non aplica o grao, que presenta tres valores para o atributo xénero, como viamos máis arriba, e que forma o plural engadindo *-s*. Agora ben, cómpre precisar que no lexicón non existen formas, estas xéranse a partir da información que

nel introducimos. Así, creamos modelos formais que reducen a flexión nas categorías variables para facilita-la introdución do vocabulario. As vantaxes deste sistema son obvias, pois subsúmense nunha única entrada, por exemplo, as 75 formas do paradigma de *cantar*, e ademais derívanse ó modelo de *cantar*, grupo V1, os cerca de 4500 verbos que seguen ese padrón. Ou sexa, seguindo con *amable*, no lexicón non figuran as formas *amable, amables,* senón que se introduce unha raíz á que se lle asigna a subetiqueta de adxectivo, grao non aplica, xénero masculino/feminino, feminino ou masculino[18], e logo remítese ó grupo de derivación G29, onde se segue un *-s* debe remata-la caracterización morfosintáctica indicando que é plural, e se non segue nada precisar que é singular.

Raíz	Subetiqueta	Grupo	Lema	Hiperlema	Normativa lema (2003)	Categoría lema	Fonte	Lexicón
amable	A0a	G29	amable	amable	si	A	volga_2004	xeral

FIGURA 5. Exemplo da entrada *amable* no lexicón de XIADA

En XIADA o lema está sempre asociado a unha clase de palabra, de maneira que se esta diverxe ou varía a súa flexión, no lexicón integraranse tantas entradas como análises diversas existan. Deste xeito hai dúas entradas para *mañá*, unha como adverbio e outra como substantivo, e outras dúas para *xerente*, pois unha só flexiona o número, análise recollida nas obras lexicográficas (DRAG, Xerais, Digalego) e a outra flexiona o xénero e o número, comportándose coma *cliente*, análise rexistrada no corpus. Ademais, como se pode observar na figura 5, cada entrada recolle tamén unha indicación sobre a normatividade do lema, a fonte da que se partiu para a súa introdución no lexicón e o módulo ó que pertence, información que pode ser de utilidade por exemplo para a construción dun corrector ortográfico, mais que non repercute de ningún modo na etiquetaxe do corpus, polo que non imos deternos nestes campos.

[18] Non debe confundirse o xeito de introduci-la información no lexicón coa información que proporciona logo a etiqueta na ferramenta de consulta. A fin de reducir esforzos para a súa introdución no lexicón, tódolos adxectivos en *-ble* caracterízanse coa subetiqueta A0a, pero logo cando se xeran as formas ou estas se anotan no corpus, ese 'a' é desagregado en tres valores diferentes: *a* (masculino/feminino), *f* (feminino) e *m* (masculino).

Por outra banda, é tamén no lexicón onde os lemas irmáns, aqueles que amosan entre si unha relación de semellanza formal inda que con pequenas variacións gráficas, se asocian a través do hiperlema, como se ve na figura 6:

Raíz	Subetiqueta	Grupo	Lema	Hiperlema	Normativa lema (2003)	Categoría lema	Fonte	Lexicón
amábe	A0a	G32	amábel	amable	si	A	volga_2004	xeral

FIGURA 6. Exemplo da entrada *amábel* no lexicón de XIADA

Grosso modo, o lema é a forma base sen os morfemas gramaticais. Así, tradicionalmente, as formas verbais agrúpanse baixo o infinitivo ou os adxectivos con oposición xenérica e numérica recóllense baixo a súa forma masculina singular. Pola súa banda, o hiperlema é un concepto relativamente novo, xerarquicamente superior, que nace para agrupar lemas que, no noso caso, se caracterizan por manifestaren entre eles unha relación de semellanza formal con pequenas variacións gráficas (Domínguez & López 2017). Variacións no mesmo sufixo (*cafetaría-cafetería*; *amable-amábel*), cambio de conxugación (*combater-combatir*, *discorrer-discurrir*), presenza de grupo culto (*dictar, construcción, rector*) ou supresión (*ditar*, *construción*) e vocalización deste (*reitor*), ou vacilacións no vocalismo (*adobo-adubo, entroido-antroido*) exemplifican algunhas das variacións ortográficas que temos en conta no establecemento do hiperlema, coincidente maioritariamente coa forma á que a normativa oficial lle concede primacía, ben explícita ou ben porque a usa nos seus textos ou é a que aparece definida no DRAG. Porén, no lexicón hai lemas que non están avalados polas autoridades académicas, mais cuxa existencia xustifica o seu emprego nos textos recollidos no CORGA. Así, nin a *plantear, plantexar* ou *prantexar* os avala ningunha autoridade, mais nos documentos concorren formas conxugadas destes lemas, polo que é forzosa a súa introdución no lemario co fin de facilita-la recuperación de información, e con esta mesma finalidade elevamos á categoría de hiperlema *plantexar*.

O lemario contido no lexicón inclúe as entradas do *Vocabulario ortográfico da Lingua Galega* (VOLG) e algo máis das 100 000 formas máis frecuentes do CORGA. Na táboa 1 poden verse os datos do lemario pertencentes, respectivamente, ó corpus de adestramento —744 530 unidades léxicas correspondentes a 31 838 lemas— e ó lexicón —1 131.282 unidades léxicas pertencentes a 59.146 lemas—, organizados por clases de palabra.

Clase de palabra	*Corpus de adestramento*	*Lexicón*
Adverbio	791	2153
Adxectivo	5254	14 717
Artigo	2	3
Conxunción	38	43
Demostrativo	3	6
Indefinido	42	41
Interrogativo-exclamativo	4	4
Interxección	94	109
Locución	527	612
Numeral	2772	1233
Posesivo	10	10
Preposición	51	60
Pronome	29	30
Relativo	5	5
Sinal de puntuación	20	20
Substantivo	18 901	33 063
Verbo	3295	7037

TÁBOA 1. Datos do lemario de XIADA

A modo de exemplo, no lexicón dáselle cabida a 6 lemas para os demostrativos, mentres que no corpus de adestramento só converxen 3. Esta diferenza nunha clase pechada que *a priori* sorprende evidencia dous feitos: i) nada garante que na selección que se realiza para servir de corpus de adestramento se rexistren tódalas etiquetas nin que concorran tódolos lemas das clases pechadas; ii) se temos en conta que os tres lemas presentes a maiores no lexicón son as variantes *iste, ise* e *aquil,* constátase unha das peculiaridades do CORGA e da súa etiquetaxe: a lingua que amosan os textos está lonxe aínda de acada-la plena estandarización, e a práctica —os textos— non sempre camiña da man do que se impón na teoría —o que prescribe a norma—, polo que o etiquetador debe facilitar, e así o fai, o recoñecemento e caracterización tanto das formas avaladas pola norma como daquelas outras que avala o uso, incluídos os dialectalismos e castelanismos.

Ata aquí tratamos unicamente as unidades léxicas simples, mais o galego caracterízase por presentar unha gran cantidade de amálgamas, polo que deseguido imos ver como se produce a análise destas.

As contraccións introdúcense nunha base de datos externa na que se recolle a seguinte información: forma gráfica da contracción, número de

compoñentes nos que se desagrega e identificación e caracterización morfosintáctica de cada compoñente, ou sexa, elemento gramatical, etiquetas que pode recibir en función da contracción na que concorre e lema ó que se remite. Entre as contraccións que están implementadas, e que polo tanto o etiquetador vai analizar desagregando nos seus elementos constitutivos outorgándolle a cada constituínte a súa etiqueta e lema, figuran tódalas formadas por preposición e artigo, demostrativo, indefinido etc., mais tamén as contraccións de pronome de dativo (*me, che, lle, nos, vos, lles*) co pronome de acusativo (*o, a, os, as, lo, la, los, las*). Así, por exemplo, malia que o pronome *me* poida caracterizarse contextualmente conforme o caso como *acusativo, dativo* ou o que denominamos *formante léxico* cando é un mero índice cos verbos pronominais, na contracción *mo* o compoñente correspondente a *me* aparece xa caracterizado no caso como *dativo*.

Porén, igual que ocorría coas unidades léxicas simples, entre as contraccións non só hai amálgamas normativas. Nos últimos anos, para facilita-lo recoñecemento automático e mellora-la recuperación de información, engadíronse numerosas amálgamas, entre elas as que provoca o uso de formas con grafías innovadoras nalgún dos compoñentes (por exemplo, *do/da, da/do, dos/das, das/dos, do/a, da/o, dos/as, das/os, d@, d@s* ou *dxs*), variantes dialectais (*dil, diles, nil, niles, diste, dise, daquil, distes, dises, daquiles, nunhos, cunhos* etc.) e mesmo formas que normativamente non se grafan xuntas (*dalí, dendeaquela, coaquela, dalgo* etc.) pero que se rexistran nos documentos do corpus. En total, 627 amálgamas.

Non obstante, a maior complexidade na análise preséntana as formas verbais cando se constrúen coa segunda forma do artigo ou levan pronomes enclíticos, posto que, a diferenza das contraccións, non estamos ante unha listaxe pechada nin o conxunto amalgamado se recolle como tal en ningunha parte do sistema XIADA. Para identificar e etiqueta-las dúas partes da unidade (*contounos, cántaa, vaite, cóme-lo, héichello*…) implementamos unha dobre conxugación para cada un dos verbos que integran o lexicón: i) por unha banda, elaborámo-la conxugación do paradigma verbal, semellante á que pode aparecer en calquera manual, só que a integramos no sistema mediante grupos formais; ii) por outro lado, construímos unha conxugación paralela formada polo paradigma conxugado completo de cada verbo presente no lexicón, pero coa particularidade de que as raíces e desinencias aí integradas son as que poden ir acompañadas dun ou máis enclíticos, incluíndo nestes a segunda forma do artigo. Así, a conxugación de i) identifica e caracteriza *contou* como forma illada no texto, mentres que a conxugación

de ii) permite caracteriza-lo *contou* inmerso en amálgamas (*contounos, contouche, contóullelo, contóuno-la*…).

Para acada-lo recoñecemento da parte dereita, a relativa ós clíticos, interveñen aínda outros recursos: i) unha táboa que contén 190 secuencias posibles de enclíticos para os que se indica, cun valor do 1 ó 4, a lonxitude que posúe a cadea; é dicir, o número de pronomes enclíticos do que consta cada secuencia; ii) unha táboa cos 36 enclíticos ou contraccións de enclíticos dunha sílaba, na que aparece o identificador da secuencia, o identificador do(s) compoñente(s), a posición de cada compoñente, o número de descomposición por se hai máis dunha análise e por último as etiquetas e lemas posibles de cada compoñente; e iii) unha serie de regras que inciden ben na segmentación e etiquetaxe dos enclíticos ben na das formas verbais cando acompañan enclíticos. Na figura 7 exemplifícase a regra segundo a cal unha forma verbal rematada en ditongo non pode concorrer cos alomorfos *o* e *lo* nin coa segunda forma do artigo como primeira parte da secuencia enclítica:

```
<rule>
  <condition>
    <target>verb_part</target>
    <content>
      <evaluation at="end">ei OR éi OR eu OR éu OR ou OR óu OR iu OR íu OR ai OR ái OR oi OR ói</evaluation>
    </content>
    <check_default>no</check_default>
    <condition>
      <target>enclitic_part</target>
      <content>
        <evaluation at="first">o OR os OR a OR as OR la OR las OR lo OR los OR -lo OR -la OR -los OR -las</evaluation>
      </content>
      <action>reject</action>
    </condition>
  </condition>
</rule>
```

FIGURA 7. Exemplo de regra

Ou sexa, estas regras restrinxen que partes esquerdas —as raíces verbais— poden combinarse con que partes dereitas —enclíticos ou combinacións de enclíticos—.

A maiores de proporciona-la análise do conglomerado, é preciso que o elemento gramatical ou unidade léxica verbal sexa o mesmo, constrúase con enclíticos ou non. Non tería sentido que nos casos de *contou* non contabilizasen os que se constrúen con dous ou máis clíticos silábicos, só porque se acentúan graficamente (*contóunolo*). Por esta razón reconstrúese a forma verbal conectando os dous lexicóns verbais: o sublexicón de clíticos identifica e caracteriza a forma verbal e logo, en función desa etiqueta e forma, o etiquetador consulta para o lema de que se trate a conxugación no lexicón principal, e reconstrúe para a forma que presenta a etiqueta coincidente coa que ten na análise e cuxa forma é máis próxima.

Ó noso parecer, o logro máis destacado da etiquetaxe do CORGA radica na consecución da distinción efectiva entre elemento gramatical e forma ortográfica, o que permite analizar nos seus elementos constitutivos, reconstruíndo a forma do elemento, a gran cantidade de contraccións e conglomerados de forma verbal e pronomes enclíticos e/ou segunda forma do artigo que presenta o galego. Con todo, non pode esquecerse que a etiquetaxe do CORGA é automática e está baseada fundamentalmente no método estatístico, polo que pode presentar erros e omisións.

Os erros poden producirse porque a unidade léxica presenta ambigüidade —ben segmental (máis dunha descomposición: *lema*), ben categorial (máis dunha etiqueta posible: *mañá*), ben atributiva (un ou máis atributos posibles: xénero de *amable*)—, e polo tanto o etiquetador non atina coa análise proposta. Por exemplo, *nos* pode se-la contracción da preposición *en* e mailo artigo, o pronome átono de primeira plural ou o alomorfo do pronome de acusativo de terceira. Pois ben, dado que a unidade *nos* só pode aparecer como elemento do paradigma do pronome de terceira se vai en énclise, seguindo o método descrito en Domínguez (2016) creámo-la regra «nos,Raa3mp,o,nos,x»[19] que lle indica ó etiquetador que a forma *nos* cando se etiqueta como Raa3mp (pronome, átono, acusativo, 3ª, masculino, plural), cuxo lema é *o*, non é unha análise posible se *nos* aparece como unidade independente; polo tanto, ó impedir esa selección, o etiquetador ten que optar por outra das análises dispoñibles: ben a de pronome de primeira plural, ben a de contracción de preposición e artigo. E aquí, de novo mediante a implementación doutra regra, corrixímo-las análises como contracción cando segue forma verbal, encarreirando o etiquetador cara á única análise posible, a de pronome de primeira plural:

en,P,en,_ os,Ddmp,o,_,x _,V*,_,_

As omisións prodúcense, como é esperable, debido a que as formas que aparecen nos textos non figuran no dicionario co que traballa o etiquetador. Dita omisión pode ser total ou parcial.

Se a omisión é total, ou sexa, se nin no lexicón nin no corpus de adestramento consta a forma en cuestión, entra en acción o módulo de adiviñación que, segundo a terminación da unidade léxica e as etiquetas dos elementos

[19] O etiquetario está dispoñible para a súa consulta nos enderezos http://corpus.cirp.gal/xiada/etiquetario/taboa> e <http://corpus.cirp.gal/xiada/etiquetario/exemplos. No primeiro preséntase unha táboa na que se recollen as clases de palabras, subtipos e valores posibles para as categorías gramaticais recoñecidas e no segundo amósase un exemplo de uso de cada etiqueta.

anteriores e posteriores, aventura unha análise. Esa análise, na maior parte dos casos atinada, non asigna lema, o que supón unha limitación á hora de recupera-la información, como describimos no apartado 5. Trátase dunha carencia que tentaremos remediar nos próximos anos construíndo un lematizador automático. Exemplifica este tipo o elemento *futboleiras* presente na secuencia *Tocou deportes e sacou a relucir as súas neuras futboleiras*.

Da outra banda, se a omisión é de tipo parcial, ou sexa, se no lexicón ou no corpus de adestramento o etiquetador atopa a forma, inda que con outra análise, non coa que corresponde ó contexto no que se localiza, vai cometer un erro sempre, pois caracteriza a unidade coa información que xa lle consta, sen cuestionala. Mostra destoutro comportamento é a unidade léxica *rulo,* acollida na versión 2.7 do lexicón[20] unicamente baixo a conxugación de *rular* como primeira persoa do presente de indicativo, de aí que tódolos demais usos de *rulo,* sexa o masculino de *rula* (*Bo foi que o meu rulo non estaba emigrado que senón xa nos quedabamos sen familia. / _¡Por santa Lilaina, que pariu por un dedo! _exclamou contento coma un rulo_.*) sexa como variante de *rolo* (*Vóltase a gradar e por fin seméntase e dase un pase de rulo.*), debido á ausencia no sistema XIADA da análise correspondente ó substantivo, vai clasificala como forma verbal de presente. Mentres que, pola contra, as ocorrencias de *rulos* vanse analizar como substantivas por mor dunha aparición con esa análise no corpus de adestramento.

Dos 44 millóns de elementos gramaticais dos que constaba a versión 3.0 do CORGA, un chisco máis de 170 000 formas ortográficas carecían de lema, ou sexa, presentaban unha omisión plena, e destas, unhas 110 000 documentábanse só unha vez. Unha análise superficial amosa a seguinte casuística: dialectalismos, castelanismos, estranxeirismos, alongamentos, erros ortográficos, erros na introdución da entrada no lexicón… Mais tamén figuran nesta relación numerosas unidades, como poden ser *futboleiro, emprendemento, privacidade, visibilizar, cousificar, rockeiro, poemario* etc., que pertencen a familias léxicas para as que están presentes outros elementos nas obras lexicográficas (*futbolista, futbolístico, emprender, emprendedor, privación, visible, visibilidade, visiblemente*…), pero que pola razón que sexa non constan aínda, o que condicionou que non se incorporasen ó sistema, pois non se detectaran. Con todo, nesta ampla listaxe non aparecen as unidades que presentan unha omisión parcial nin temos xeito de extraelas, máis ca coa observación directa dos resultados.

[20] Dispoñible no enderezo http://corpus.cirp.gal/xiada/descargas/texto_lexico dende o 1/4/2019.

A curto e longo prazo o camiño a seguir é mellora-la etiquetaxe, e para iso temos que actuar en dúas frontes: completando o lexicón e corrixindo análises que son erróneas.

A primeira actuación abrangue a introdución no lexicón de novas entradas, para o que partiremos da listaxe de formas sen lema, ou posteriores actualizacións, e teremos en conta a súa frecuencia de uso, co fin de incidir sobre o maior número posible de elementos. Así mesmo, implementaremos tamén no lexicón as desinencias que posibiliten a análise morfosintáctica das formas verbais, soas ou combinadas con enclíticos, rematadas en ditongo decrecente acentuado (*seréi, teréi, léuse, estóuche*...) [21], as que ofrecen a variante *-che* para a segunda singular do pretérito de indicativo (*deche, fixeche, viñeche*...), as que amosan *-chedes* para a de plural (*apostáchedes, botáchedes, déchedes, víchedes*...), a vogal temática *-e* para a terceira singular do pretérito dos verbos da terceira conxugación (*abreu, escribeu, parteu*), a variante con *-ñ* na raíz do presente de subxuntivo de *ser* (*seña, señas, señamos*...) etc.

Respecto da segunda actuación, a inxente cantidade de amálgamas de forma verbal e pronome enclítico e/ou segunda forma do artigo que posúe o galego levan a que, en casos de ambigüidade segmental, ou sexa cando unha unidade é descompoñible de máis dun modo, os datos estatísticos provocan que o etiquetador opte a miúdo pola análise desagregada. É o que ocorre por exemplo con *lema,* que presenta dúas alternativas de segmentación:

1. Elemento gramatical *lema*. Non hai segmentación, ou sexa, o elemento gramatical e a unidade son coincidentes, polo que corresponde a análise como substantivo común masculino singular, calquera que sexa a acepción na que se inscriba.
2. A unidade ortográfica descomponse en tres elementos gramaticais: *le, me* e mais *a*. O verbo ademais pode caracterizarse como forma imperativa de segunda singular ou indicativa de terceira singular, mentres que *me* só pode ser pronome de dativo por ir contracto co pronome de acusativo *a*.

Pois ben, para actuar contra a análise que apoia a estatística baseada só no peso das etiquetas, o método probabilístico debe apoiarse na elaboración de regras lingüísticas que guíen o etiquetador nestes casos en que a estatís-

[21] Estas terminacións implementámolas xa no sistema XIADA para os verbos regulares e para os irregulares que seguen o modelo *sacar, cocer* ou *traducir,* así como a terminación *-che* para a segunda singular do pretérito, polo que xa son efectivas na versión 3.1 do CORGA. Ou sexa, ante unha consulta polo lema *sacar* obterémo-los casos tamén de *saquéi*, *sacóu, sacóuno* ou *sacache*. Porén falta introducilas nos demais modelos irregulares.

tica erra sistematicamente, e para iso examinarémo-las construcións nas que participan ámbalas estruturas e tentaremos establece-los usos do substantivo impedindo a análise desagregada se o precede o artigo, posesivo ou demostrativo, por exemplo. Mais esta é só unha mostra no nivel segmental; existen outras moitas actuacións no nivel categorial sobre as que cómpre tamén incidir, como pode ser favorece-la distinción entre *a* preposición e artigo ou fixar cando a paréntese de apertura ou de peche é sinal de puntuación e cando en realidade forma parte dunha unidade léxica que, como amosan as seguintes dúas concorrencias extraídas do CORGA, todos entendemos cando as visualizamos, mais que resulta complexo que procese unha ferramenta automática:

(2) O(s) tempo(s) está(n) louco(s)

(3) Se por unha vez, con esta película, Warhol se achega ao cinema de montaxe ou de material preexistente, con esta repetición do plano afianza ese parentesco co(a)s cineastas estruturais, aos que antecede, que verán na montaxe en bucle un fértil recurso estilístico.

Por outra banda, sabemos que un foco importante de erro na etiquetaxe concéntrase nos nomes propios, para os que neste momento funciona unicamente un módulo de recoñecemento nos textos de ficción. Segundo este módulo, o etiquetador comproba se a unidade que se documenta en posición inicial se rexistra tamén en interior de secuencia, e etiqueta de acordo con esta información. Ou sexa, ante un *Rosa* situado a comezo de enunciado —posición ambigua, posto que pode ser nome propio ou corresponder a un elemento do léxico común que se grafa con maiúscula inicial por convención—, o etiquetador verifica se concorren outros casos de *Rosa* en situación medial —posición non ambigua—. Se rexistra algún, etiqueta o primeiro como propio e se non trátao como nome común. Porén, aplicar este procedemento a un xornal ou revista considerado globalmente non deu os froitos esperados, polo que queremos estudar outras posibilidades, entre as que figura testa-lo procedemento anterior aplicándoo noticia a noticia ou crear un lexicón de propios.

Por último, en relación cunha posible subclasificación das entidades, barallamos tamén a posibilidade de introducir no lexicón información de tipo semántico, dado que por exemplo os trazos + ou – animado clasificando os substantivos sería de enorme axuda para afronta-lo estudo do emprego da preposición *a* nos complementos directos, o mesmo que o trazo + humano devén fundamental para ver que substantivos presentan formas diferenciadas para o xénero feminino e o masculino.

5. A APLICACIÓN DE CONSULTA

Para facilita-la recuperación de información que ofrece o CORGA deseñouse una aplicación de consulta, dispoñible en http://corpus.cirp.gal/corga, que se organiza en varias pestanas, das cales as máis salientables son *Información, Guía, Frecuencias* e *Buscas*, esta última a cerna da aplicación (figura 8).

FIGURA 8. Pantalla inicial de captación de datos

Os datos presentes en *Contacto* e *Equipo* son transparentes. *Contacto* dispón dun formulario para comunicar erros de funcionamento ou solicitar axuda puntual sobre algunha consulta, mentres que *Equipo* recolle a relación de persoas que participan actualmente no proxecto ou que colaboraron nalgún momento.

Na parte esquerda, *Información*, tal e como o seu nome indica, acolle diversa información de utilidade para o usuario, entre a que cómpre salienta-los criterios que se empregan para a selección dos documentos, a relación de documentos contidos no corpus, as frecuencias tendo en conta tanto os criterios de selección (*data, tipo de texto* e *área temática*) coma os de clasificación (*orixe, bloque, xénero* e *subtipo de documento*), o historial de versións, a listaxe de preguntas máis frecuentes ou as publicacións que empregaron o corpus como fonte para o seu traballo.

Pola súa banda, a pestana *Guía* acolle un manual de uso detallado do sistema de recuperación de información, así como a descrición xeral de como se estruturan e codifican os textos, a relación das marcas e mailo etiquetario que empregamos, respectivamente, para a súa codificación e anotación morfosintáctica.

As marcas de codificación, ofrecidas cunha pequena descrición e un exemplo de aplicación, reflíctense nos resultados das buscas amosando as

palabras afectadas sobre fondo amarelo ou branco. Basta situa-lo rato enriba para que a cor se traduza no texto que indica a marca concreta da que se trata e cuxos valores vimos polo miúdo no apartado *3.1. Marcas de codificación*. A cor amarela ofrece información sobre características lingüisticamente relevantes (no rexistro oral as marcas máis habituais corresponden a *alongamento* e *palabra cortada,* mentres que no rexistro escrito a marca máis frecuente é *outra lingua,* seguida de *texto en verso* ou *texto en táboa*), mentres que a cor branca sitúa o texto nunha acoutación ou asígnao a un interlocutor específico e só é visible dende o contexto.

O etiquetario empregado na caracterización morfosintáctica do corpus, recollido tamén baixo a pestana *Guía* e de recomendable consulta para a formulación das procuras de tipo gramatical, ofrécese nunha dobre vertente: i) sintética: nunha táboa recóllense as clases de palabras recoñecidas, os atributos que se consideran pertinentes para cada unha delas e a posición que cada atributo ocupa na cadea resultante; ii) analítica: organizada por clases de palabra, recóllese para cada etiqueta o seu significado e un exemplo de uso. Así mesmo, conscientes da dificultade de que os usuarios manexen un etiquetario complexo, na ferramenta de consultas deseñouse un menú amigable para a introdución da etiqueta, de xeito que non hai que coñece-lo sistema empregado para formular unha procura gramatical. Así, se lle prouguer, o usuario accede a un menú que, como se observa na figura 9, o vai guiando nas alternativas dispoñibles, primeiro segundo a clase de palabra considerada e, logo, segundo os subtipos e valores categoriais pertinentes ou posibles en función das eleccións previas.

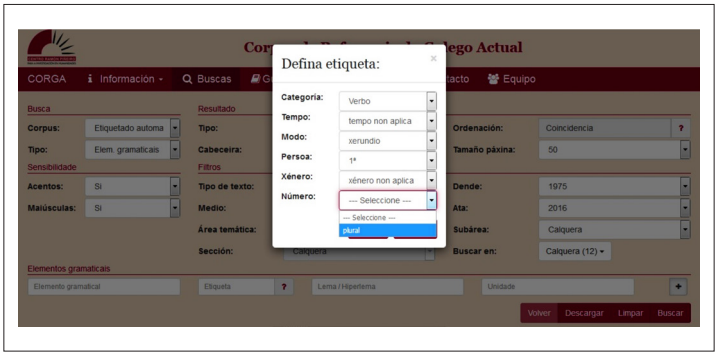


FIGURA 9. Exemplo de definición de etiqueta a través do menú amigable

Así mesmo, na observación dos resultados, for nas concordancias for no contexto destas, sempre que estea explícita unha etiqueta, a ferramenta desenvólvea con só situar enriba o punteiro do rato, pois ó igual que ocorre coas marcas de codificación emerxe un cadro de texto que a desagrega proporcionando a caracterización morfosintáctica da unidade en cuestión nunha lingua comprensible para calquera usuario. Esta dobre actuación, menú despregable para formula-las consultas e cadro de texto emerxente co desenvolvemento da etiqueta nas concordancias ou na análise completa da secuencia visible dende o contexto, permite que un usuario non familiarizado co sistema de etiquetaxe aplicado poida realizar consultas a través da modalidade *elementos gramaticais* e comprender facilmente os resultados.

Pola súa parte, baixo a pestana *Frecuencias* atópanse as listaxes de formas, elementos gramaticais, lemas, hiperlemas e etiquetas desagregadas segundo correspondan ó CORGA xeral —alternativa denominada *corpus etiquetado automaticamente* no ítem *Corpus* do bloque *Busca* no sistema de consultas— ou ó subcorpus desambiguado á man que serviu de adestramento para o etiquetador automático —opción designada *corpus etiquetado manualmente*—. Os arquivos situados debaixo de mil e cinco mil máis frecuentes poden consultarse directamente dende a aplicación, mentres que para visualiza-los listados completos —clasificados por *palabras ortográficas, elementos gramaticais, lemas, hiperlemas* e *etiquetas—,* debido ó seu peso, cómpre descarga-los arquivos correspondentes.

Por fin, o sistema de consultas propiamente dito, a única parte dinámica e cerna da aplicación, ábrese na pestana *Buscas* e permite *grosso modo*:

— realizar consultas léxicas e gramaticais combinándoas con criterios de clasificación textual (data, tipo de documento, área temática, medio, xénero etc.)
— reproduci-lo fragmento de audio que corresponde ó resultado da consulta
— descarga-los resultados no formato tsv (valores separados por tabulacións)

As consultas poden realizarse sobre a totalidade do corpus, opción por defecto que amosa a figura 8, ou ben sobre o subcorpus desambiguado á man, o que serviu de adestramento para o etiquetador, incluído baixo o ítem *Corpus* coa denominación de *corpus etiquetado manualmente*. Ademais, é posible referi-las procuras á totalidade do texto, que é a opción habitual, ou ben, mediante a selección dunha das alternativas presentes no bloque *Buscar*

en, centralas en fragmentos que se localizan nunha parte do texto específica (*apéndice, cita, corpo, dedicatoria, encabezamento, nota, pé de foto, prólogo, resumo* ou *titular*) ou que remiten o texto a unha parte dialogada na que o emisor estaba marcado explicitamente (*interlocutor*) ou a unha didascalia dramática (*acoutación*).

A parte do texto na que se sitúa o obxecto da busca, ademais de ser parámetro de consulta a través da opción *Buscar en*, ocupa a última posición na información que se fornece como metadatos na cabeceira de cada concordancia: *título, autor, editorial* ou *url, ano, medio, orixe, bloque, xénero, subtipo* e *sección no texto*.

O usuario pode tamén crear subcorpus virtuais a partir das escollas que realice entre os parámetros de clasificación dispoñibles, cuxas denominacións e valores enumeramos deseguido:

— ***Dende***... ***Ata***...: dende 1975 ata 2016, primeiro e último valor respectivamente, pode seleccionarse calquera rango de anos.
— **Tipo de texto**. Comprende á súa vez catro parámetros cuxos valores se activan ou desactivan en función das escollas previas (a selección de *ficción* determina que no xénero os valores só poidan ser *calquera, dramático* ou *narrativo* e no *Subtipo calquera, novela, relato curto, obra de teatro* ou *guión*):
  — Orixe: *Calquera, Escrita, Oral*
  — Bloque: *Calquera, Ficción, Non ficción*
  — Xénero: *Calquera, Dramático, Ensaístico, Narrativo, Xornalístico*
  — Subtipo: *Calquera, Novela, Relato curto, Obra de teatro, Guión, Libro de texto, Memoria, Artigo científico, Divulgación, Xornal, Revista, Blog, Tertulia, Informativo, Variedades, Programa cultural, Publicidade*
— **Medio**: *Calquera, Xornal, Revista, Libro, Internet, Audiovisual*
— **Área temática** e **Subárea**. Véxase a figura 2 do apartado *2. Deseño e composición*, onde xa foron presentados os valores posibles para este parámetro. É aplicable unicamente para a procedencia *escrita*, o bloque da *non ficción*, o xénero *ensaístico* e *xornalístico* e os subtipos de documento *Libro de texto, Memoria, Artigo científico, Divulgación, Xornal, Revista* e *Blog*.
— **Sección.** Parámetro só aplicable ás noticias procedentes de xornais, para as que a agrupación en seccións é natural. As posibilidades son: *Actualidade, Área de Compostela, Campus, Cultura e Sociedade, Deportes, Economía, España, Galicia, Internacional, Opinión, Suplemento, Tempo* e *TV*.

Por outra parte, o parámetro referido á *Sensibilidade* permite que na recuperación de información se teñan en conta as diferenzas debidas ó emprego de acentos e maiúsculas, de xeito que o usuario decide se para a súa consulta é relevante a distinción entre formas con tiles e formas sen tiles e/ou grafadas en minúsculas ou maiúsculas ou, pola contra, desexa que se ignoren esas diferenzas. Trátase sen dúbida dunha funcionalidade moi útil para calquera corpus, pero máis se cabe no caso que nos ocupa ante a enorme variación que presentan os textos e que evidencian o proceso de normativización dunha lingua minorizada.

O usuario dispón así mesmo doutras alternativas para refina-las buscas, ben mediante comodíns ben mediante operadores booleanos. Con respecto ós comodíns, o signo de peche de interrogación (?) e o asterisco (*) conmutan caracteres. O sinal de interrogación de peche substitúe un carácter, de xeito que a consulta por *ga?ela* devolve tanto os casos de *gabela*, coma os de *gavela*, *gacela* ou *gamela*. O asterisco substitúe un carácter, varios ou ningún, polo que a procura **mente* devolverá tódolos adverbios rematados en *-mente*, mais tamén calquera outra palabra que conteña ese segmento final, coma os substantivos *mente* ou *semente,* os adxectivos *clemente* ou *demente* e as formas verbais *alimente* ou *argumente*, entre outras.

Os dous operadores booleanos que están activos no sistema de consulta son os equivalentes a OU e NON. O signo de peche de admiración (!) equivale a NON e a barra vertical (|) a OU, de xeito que coa consulta *ga?ela*|*ga?elas!gacela*!gamela** recupéranse tódolos casos de *gabela* ou *gavela,* singulares e plurais, e prescíndese do ruído que introducirían os resultados de *gamela* e *gacela,* singulares e plurais.

Calquera dos parámetros clasificatorios que vimos máis arriba ou das posibilidades de refina-la procura habilitando a sensibilidade a tiles e maiúsculas e/ou empregando comodíns e mais operadores booleanos poden cruzarse cos distintos tipos de consulta textual que permite levar a cabo a aplicación e nos que nos imos centrar a continuación.

A procura textual, establecida por defecto na aplicación para unha consulta de tipo léxico, organízase no bloque *Busca* a través de *Tipo* cara á consulta mediante a modalidade de *palabras ortográficas* ou cara á modalidade de *elementos gramaticais*. Isto permite, *grosso modo,* buscar:

— Palabras ortográficas*: cancela, falar, disllo, nel, ao redor do…*
— Elementos gramaticais: *cancela, falar, dis* (incluíndo tódolos casos con pronomes enclíticos nos que entra *dis*: *dilo, disllo, dísnolo* etc.), *el*

(incluíndo as contraccións *nel* e *del*), *ao redor de* (incluíndo os casos de *ao redor desta, ao redor duns, ao redor das etc.*).

— Lemas: *cancela* (inclúe os casos de *cancela* e *cancelas*), *falar* (tódalas formas do paradigma do verbo *falar*), *el* (tódolos casos do pronome de terceira, masculinos ou femininos, singulares ou plurais, e mailas concorrencias da forma arcaica do artigo determinado), *alumno* (inclúe *alumno, alumna, alumnos, alumnas, alumn@, alumn@s, alumno/a, alumna/o, alumnos/as, alumnas/os, alumno/alumna, alumnos/alumnas*).
— Hiperlemas: *ata* (agrupa os lemas *ata, até, ate, hasta* e *hastra*), *ditar* (agrupa os lemas *ditar* e mais *dictar*, polo que devolve tódalas ocorrencias con calquera das formas verbais de calquera deses dous lemas).
— Clases de palabras: *substantivo,* por exemplo.
— Valores das subcategorías gramaticais aplicables en cada caso: *grao, xénero* e *número* no caso dos adxectivos, por exemplo.

A consulta básica é a que corresponde á modalidade de *palabras ortográficas*, soportada pola forma gráfica, coa que poden recuperarse ata un máximo de 5 palabras ou segmentos sucesivos. Esta modalidade é de utilidade cando a cadea que se procura é coincidente coa grafía e non entra en xogo a flexión nominal nin verbal ou calquera outro tipo de información gramatical. Así, a consulta textual *aí a vén* devolve os casos nos que concorre esta expresión literal, verdadeiramente útil se o comparamos co que suporía realiza-la consulta manual nos textos, mais paupérrima se temos en conta que o pronome de acusativo pode variar en xénero e número e que o verbo pode aparecer noutro número ou tempo verbal. Para supera-las limitacións que impoñen as consultas fundamentadas en formas ortográficas concretas creouse a modalidade de consulta por *elementos gramaticais*, a cal, grazas á etiquetaxe do corpus, permite dar un salto cualitativo na recuperación de información e a posterior análise gramatical ó poder introducir nos parámetros de consulta información de carácter gramatical, non só xa a forma léxica, senón tamén lemas (tódalas formas do verbo *vir*), clases de palabras (adverbio, por exemplo) ou valores das subcategorías gramaticais aplicables en cada clase de palabra (tonicidade, caso, xénero e número para o pronome persoal, por exemplo).

O seguinte caso práctico evidencia tanto a potencialidade que ofrece o motor de buscas do CORGA como a importancia que adquire para a obtención de datos lingüísticos e posterior análise destes a posibilidade de introducir información gramatical na formulación das buscas. Imos comproba-la vitalidade dunha estrutura tipicamente galega que constitúe unha anomalía

para a sintaxe: as estruturas focalizadas de adverbio de lugar con pronomes de acusativo en construcións intransitivas, tipo *aí a vén*. Para iso, selecciónase a procura por elementos gramaticais e créanse 3 liñas para a análise de 3 elementos sucesivos. Como mostra a figura 10, introducimos no lema do primeiro elemento os adverbios *aquí|aí|alí|velaí|velaquí|alá|acolá* botando man do carácter que converte a busca en expresión regular, a barra vertical que equivale a un OU, de xeito que poida aparecer calquera deles. No segundo elemento recorremos ó menú de introdución de etiquetas e seleccionámo-los trazos correspondentes a Pronome átono acusativo de terceira (Raa3*). Por último, no terceiro elemento introducimos na caixa do lema os verbos que con máis frecuencia aparecen nesta estrutura: *ir|vir|andar|estar|quedar* e prememos en *Buscar*.

FIGURA 10. Pantalla de captación de datos para recuperar casos da estrutura tipo *aí a vén*

Con estes datos obtemos:
1) As concordancias (figura 11) no formato KWIC (*Key Word in Context*)

FIGURA 11. Pantalla coas concordancias iniciais

Os resultados, ofrecidos por defecto ordenados pola coincidencia, poden organizarse por un ou varios dos seguintes parámetros: *área temática, coincidencia, data, documento, etiqueta, lema, medio, palabra anterior, palabra posterior, segunda palabra anterior, segunda palabra posterior, compoñente 2, compoñente 3, compoñente 4, compoñente 5, etiqueta 2, etiqueta 3, etiqueta 4, etiqueta 5, lema 2, lema 3, lema 4* e *lema 5*.

Así mesmo, os resultados inclúen a posibilidade de descarga-las concordancias, reproduci-lo audio no caso de se localizaren en transcricións, e finalmente de acceder a unha ampliación do contexto onde ademais se recolle a análise morfosintáctica representada en 3 liñas: a primeira para os elementos gramaticais, a segunda para as etiquetas morfosintácticas correspondentes a cada un dos elementos da fila superior, e por último, a terceira, o lema/hiperlema ó que se remite cada elemento. A figura 12 dá conta desta representación:

FIGURA 12. Visualización do contexto e a análise morfosintáctica

2) A frecuencia simple (figura 13) modificando na opción *Tipo* do bloque *Resultado*:

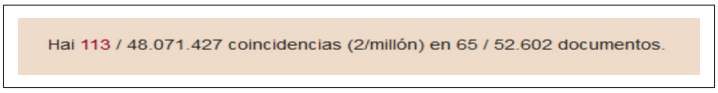


FIGURA 13. Información sobre a frecuencia simple

3) As frecuencias completas segundo os diversos parámetros de clasificación do corpus. Estes datos fanse visibles a través de dúas modalidades de representación diferentes: i) os datos dispóñense a xeito de táboas por cada un dos parámetros de clasificación do corpus (figura 14); e ii) os datos tras-

ládanse a gráficas para proporcionar unha visualización máis cómoda e facilita-la obtención dunha idea xeral rápida de como se distribúen (figura 15)[22]:

Lustro

	Coincidencias	Documentos	Frec. norm.
1975-1979	1 / 964.816	1 / 117	1/millón
1980-1984	2 / 1.674.613	1 / 143	1/millón
1985-1989	1 / 2.020.451	1 / 120	0/millón
1990-1994	7 / 5.419.144	6 / 386	1/millón
1995-1999	49 / 9.889.325	21 / 12.423	5/millón
2000-2004	12 / 7.388.909	9 / 9.479	2/millón
2005-2009	27 / 9.566.361	16 / 13.877	3/millón
2010-2014	14 / 10.300.109	10 / 14.478	1/millón
2015-2019	0 / 960.284	0 / 1.627	0/millón

Área temática

	Coincidencias	Documentos	Frec. norm.
Economía e política	2 / 14.813.896	2 / 26.966	0/millón
Cultura e artes	0 / 5.113.154	0 / 8.980	0/millón
Ciencias sociais	4 / 12.229.792	4 / 12.098	0/millón
Ciencias e tecnoloxía	1 / 6.240.077	1 / 8.734	0/millón
Outros	0 / 8.972.767	0 / 21.411	0/millón

Medio

	Coincidencias	Documentos	Frec. norm.
Xornal	2 / 14.571.997	2 / 41.122	0/millón
Revista	0 / 6.607.736	0 / 8.502	0/millón
Libro	109 / 26.556.176	62 / 2.809	4/millón
Internet	0 / 80.425	0 / 160	0/millón
Audiovisual	2 / 367.678	1 / 57	5/millón

Orixe

	Coincidencias	Documentos	Frec. norm.
Oral	0 / 266.322	0 / 50	0/millón
Escrita	113 / 47.915.690	65 / 52.600	2/millón

Bloque

	Coincidencias	Documentos	Frec. norm.
Ficción	109 / 17.644.668	61 / 2.160	6/millón
Non ficción	4 / 30.271.022	4 / 50.440	0/millón

Suptipo

	Coincidencias	Documentos	Frec. norm.
Novela	72 / 12.243.468	30 / 203	6/millón
Relato curto	18 / 3.553.671	16 / 1.776	5/millón
Obra de teatro	17 / 1.739.867	14 / 173	10/millón
Guión	2 / 107.642	1 / 8	19/millón
Libro de texto	0 / 634.422	0 / 7	0/millón
Memoria	0 / 575.944	0 / 10	0/millón
Artigo científico	0 / 378.631	0 / 52	0/millón
Divulgación	2 / 7.671.952	2 / 617	0/millón
Xornal	2 / 14.571.997	2 / 41.122	0/millón
Revista	0 / 6.357.651	0 / 8.472	0/millón
Blog	0 / 80.425	0 / 160	0/millón
Tertulia	0 / 69.310	0 / 9	0/millón
Informativo	0 / 63.809	0 / 11	0/millón
Variedades	0 / 109.298	0 / 16	0/millón
Programa cultural	0 / 24.513	0 / 4	0/millón
Publicidade	0 / 1.392	0 / 10	0/millón

Xénero

	Coincidencias	Documentos	Frec. norm.
Narrativo	90 / 15.797.159	46 / 1.979	6/millón
Dramático	19 / 1.847.509	15 / 181	10/millón
Ensaístico	2 / 9.260.949	2 / 686	0/millón
Xornalístico	2 / 21.010.073	2 / 49.754	0/millón

FIGURA 14. Datos das frecuencias completas

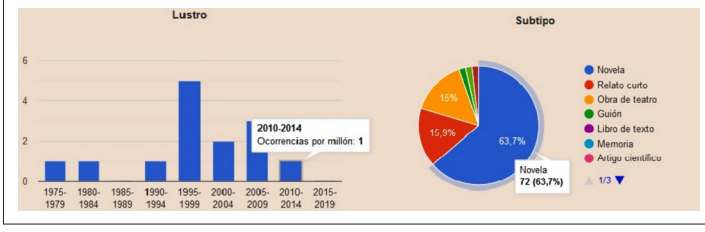


FIGURA 15. Gráficas coa frecuencia normalizada por lustros e subtipo de documento

Como podemos ver, trátase dunha construción relativamente usual no corpus, con 113 casos, en singular e plural, co verbo en presente, mais tamén

[22] Por motivos de espazo amosamos unicamente a gráfica temporal e a de subtipo de documentos.

en pasado, representada en tódolos lustros, empregada maioritariamente na ficción e nos xéneros narrativo e dramático, ausente nas transcricións, mais isto pode deberse ó tipo de programas seleccionados etc. Ou sexa, o que queda é analizar en profundidade os datos que fornece o corpus.

Hai, non obstante, polo momento unha limitación nas consultas por lema ou hiperlema que convén ter presente: se no lexicón non consta o lema, o etiquetador, a través do módulo de adiviñación, aventura unha etiqueta tendo en conta a terminación da propia palabra descoñecida e das análises que constitúen o seu contexto, pero non ofrece lema. Isto tradúcese na aplicación de consulta do seguinte xeito:

1. Se se realiza unha consulta por un lema ou hiperlema que non está no dicionario/lexicón de XIADA, o sistema indicará que non atopa resultados, inda que o corpus rexistre ocorrencias dalgunha forma do paradigma dese lema ou hiperlema.
2. Se se realiza a consulta por etiqueta ou elemento gramatical e se accede á análise completa da secuencia a través do contexto, na terceira liña, a correspondente ós lemas, aparecerá para os elementos descoñecidos un asterisco.

A posibilidade de utilizar directamente trazos gramaticais, independentes das formas que os soportan, ilustra a potencia da recuperación de información e a súa utilidade para a análise gramatical, pero o sistema de explotación do CORGA conta ademais con outra opción que permite supera-las restricións que impón a linearidade, permitindo as consultas por proximidade. Ofrécese así a opción de realizar consultas, tanto na modalidade de palabras ortográficas (dúas palabras ou dous segmentos destas) coma de elementos gramaticais (dous lemas, dous hiperlemas, dúas etiquetas, dous elementos gramaticais ou calquera combinación destes), nas que o elemento 1 —e aquí elemento é calquera das opcións que acabamos de mencionar— está con respecto ó elemento 2 á distancia que especifique o usuario —un valor entre 1 e 10, exacto ou inferior ó indicado—. Esta alternativa permite que se interpolen unidades entre os dous elementos introducidos, co que, a modo de exemplo, procurando o hiperlema /*poñer* —información presente no Elemento 1— que concorra co lema *verde* —información achegada para o Elemento 2— a unha distancia máxima de 4 ou menos elementos recuperaremos tódolos casos de *poñer* ou *pór verde a alguén* (*Pepe* ponnos verdes*, que se somos uns pequenos burgueses* [...]; [...] *xuramos escribir a Lonely Planet* póndoos verdes; [...] *tomou cartas no asunto para* poñer

verde *á esposa de Clinton*; *aproveitades calquera ocasión para* poñerme verde *ás miñas costas* [...] etc., con independencia de que existan casos de énclise pronominal ou interpolación doutros elementos. En contrapartida, recupéranse tamén resultados que non corresponden ó que queriamos buscar, inda que si cumpren cos criterios de busca, como son, entre outros: *ergueuse,* puxo *a bata* verde *e calzou as zapatillas* ou pon*se en* verde *o semáforo*. Agora ben, canto maior sexa a fixación da estrutura que se busca, menos ruído se introducirá nos resultados.

Para rematar, cómpre facer fincapé na flexibilidade e potencialidade da ferramenta que permite empregar nunha mesma consulta comodíns, operadores booleanos, sensibilidade a acentos ou maiúsculas e variables clasificatorias dos documentos combinándoos cos distintos tipos de modalidade de busca, por palabras ortográficas ou elementos gramaticais, ben sucesivos ben descontinuos, o que converte o CORGA nunha ferramenta moi útil para obter datos da lingua galega actual de tipo léxico, gramatical, terminolóxico, fraseolóxico, discursivo etc.

Recursos electrónicos

CORGA: Centro Ramón Piñeiro para a investigación en humanidades: *Corpus de Referencia do Galego Actual (CORGA) [3.1]* http://corpus.cirp.gal/corga/

DIGALEGO: *Dicionario de galego* (Ir Indo) https://digalego.xunta.gal/digalego

DRAG: *Dicionario da Real Academia Galega* https://academia.gal/dicionario

ELAN: Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands https://tla.mpi.nl/tools/tla-tools/elan/

EAGLES: Eagles (1996): Recommendations for the morphosyntactic annotation of corpora. EAGLES Document EAG-TCWG-MAC/R. http://www.ilc.cnr.it/EAGLES96/browse.html

TextPad: The editor for Windows https://www.textpad.com/

VOLG: *Vocabulario ortográfico da lingua galega* https://academia.gal/recursos-volg

XIADA: Centro Ramón Piñeiro para a investigación en humanidades: *Etiquetador/Lematizador do Galego Actual* [2.7] http://corpus.cirp.gal/xiada

XMLmind: XMLmind XML Editor http://www.xmlmind.com/xmleditor/

Referencias bibliográficas

Barcala Rodríguez, Fco. Mario (2010): *Corpus lingüísticos estruturados de grandes dimensións: Metodoloxía e sistemas de recuperación de información. Tese de doutoramento*. Universidade da Coruña. http://hdl.handle.net/2183/7171

Caíña Hurtado, María, Eva María Domínguez Noya & María Sol López Martínez (2019): «A linguaxe non sexista no CORGA: descrición e reflexión sobre as variantes empregadas». *Madrygal. Revista de Estudios Gallegos* 22, 73-91. https://doi.org/10.5209/madr.66853

Domínguez Noya, Eva, Fco. Mario Barcala & Miguel Ángel Molinero (2009): «Avaliación dun etiquetador automático estatístico para o galego actual: Xiada», *Cadernos de Lingua* 30-31, pp. 151-193.

Domínguez Noya, Eva María (2013): *Etiquetaxe e desambiguación automáticas en galego: o sistema XIADA*, Tese de doutoramento. Universidade de Santiago de Compostela. http://hdl.handle.net/10347/9587

Domínguez Noya, Eva María (2016): «O etiquetador probabilístico de XIADA e o seu teito de acerto: a elaboración de regras lingüísticas», in Manuel González González (ed.): *Lingua, pobo e terra: estudos en homenaxe a Xesús Ferro Ruibal*, Santiago de Compostela: Xunta de Galicia / Centro Ramón Piñeiro para a investigación en humanidades, pp. 213-232.

Domínguez Noya, Eva María & María Sol López Martínez (2017): «*Tratamento da variación lingüística no CORGA*», *in Marta Negro Romero, Rosario Álvarez & Eduardo* Moscoso Mato (eds.): *Gallaecia: estudos de lingüística portuguesa e galega*, Santiago de Compostela: Universidade de Santiago de Compostela, pp. 421-440. https://doi.org/10.15304/cc.2017.1080.22

Domínguez Noya, Eva María & Fco. Mario Barcala Rodríguez (2018): «Grafías innovadoras na linguaxe non sexista: unha proposta para a súa etiquetaxe automática», in Marta Díaz, Gael Vaamonde, Ana Varela, Mª Carmen Cabeza, José M. García-Miguel & Fernando Ramallo (eds.): *Actas do XIII Congreso Internacional de Lingüística Xeral*. Vigo: Universidade de Vigo, pp. 291-298.

Rojo, Guillermo, Marisol López Martínez, Eva María Domínguez Noya & Fco. Mario Barcala (2016): «O corpus de referencia do galego actual (CORGA): estado actual e perspectivas», in Manuel González González (ed.): *Lingua, pobo e terra: estudos en homenaxe a Xesús Ferro Ruibal*, Santiago de Compostela: Xunta de Galicia / Centro Ramón Piñeiro para a investigación en humanidades, pp. 445-473.

CORILGA: UN CORPUS PARA O ESTUDO DA VARIACIÓN E DO CAMBIO LINGÜÍSTICO NO GALEGO FALADO

CORILGA: a corpus for the study of variation and linguistic change in spoken Galician

Elisa Fernández Rei
Xosé Luís Regueira
Instituto da Lingua Galega, Universidade de Santiago de Compostela

Resumo

A necesidade de impulsar os traballos sobre variación e cambio lingüístico no galego actual, e de maneira especial os orientados ao estudo da lingua falada, levou a que no seo do Instituto da Lingua Galega (ILG) da Universidade de Santiago de Compostela se puxese en marcha a construción dun corpus oral que contase con transcricións aliñadas coas gravacións de audio: CORILGA (Corpus Oral Informatizado da Lingua Galega). Este corpus incorpora un número aínda reducido de textos, mais xa cobre un importante rango tanto de variación lingüística como de espazo temporal.

Ademais, o corpus incorpora unha serie de ferramentas de axuda á transcrición e á etiquetaxe dos datos: un aliñador automático texto-voz, unha ferramenta de recoñecemento de fala e de transcrición automática, un lematizador e tamén a adaptación ao galego do etiquetador morfolóxico Freeling.

O seu sistema de buscas permite levar a cabo un filtrado dos datos, de xeito que se poden comparar resultados en textos de diferentes tipos e graos de formalidade ou segundo as características sociolingüísticas dos informantes para o estudo da variación. Así mesmo, utilizando os filtros de data da gravación e do tramo de idade dos informantes o CORILGA permite tamén o estudo do cambio lingüístico tanto en tempo aparente como en tempo real.

Palabras chave: corpus de lingua falada, variación lingüística, cambio lingüístico, tecnoloxías da fala

Abstract

The need to promote studies on variation and linguistic change in present day Galician, and especially those aimed at the study of the spoken language, led the Institute of Galician Language (ILG) of the University of Santiago de Compostela to start the construction of a spoken corpus with transcripts aligned with audio recordings: CORILGA (Computerised Spoken Corpus of the Galician Language). As yet, this corpus incorporates a reduced number of texts, but it already covers an important range of both linguistic variation and time (more than 40 years).

In addition, the corpus incorporates a series of tools to enable transcription and data labelling: an automatic text-to-speech alignment tool, a speech recognition and automatic transcription tool, a lemmatiser and also the adaptation to Galician of the Freeling morphological labelling tool.

Its search system enables data filtering, so that results can be compared in texts of different types and degrees of formality or according to the sociolinguistic characteristics of the informants for the study of variation. At the same time, using the data filters of the recording and the age range of the informants, CORILGA also enables the study of linguistic change in both apparent and real time.

Keywords: corpus of spoken language, linguistic variation, linguistic change, speech technologies

1. Introdución

A necesidade de asentar os estudos lingüísticos en datos provenientes das producións e interaccións reais, ou o máis próximas a elas que sexa posible, deu lugar a un cambio teórico e metodolóxico moi profundo na lingüística de mediados do século pasado: a reflexión lingüística deixou de basearse na introspección e na competencia lingüística de quen levaba a cabo a investigación para pasar a fundamentarse en datos extraídos de contextos reais e avaliados como representativos das producións cotiás dos falantes. Son exemplos paradigmáticos deste cambio de modelo a aparición de disciplinas e orientacións teóricas como a sociolingüística variacionista (Labov 1966, 1972) ou a lingüística do texto, a análise da conversa (fundamentalmente a partir de Sacks *et al*. 1974), a etnografía da comunicación (Gumperz e Hymes 1964, 1972) ou a sociolingüística interaccional (Gumperz 1982). Este cambio de orientación na lingüística supuxo a súa conversión nunha ciencia empírica, onde os córpora cobraron cada vez unha maior relevancia para o estudo das diferentes compoñentes lingüísticas e en moitas das súas orientacións ou perspectivas: estudos fonéticos e fonolóxicos ou de corte gramatical, traballos desde unha perspectiva pragmática ou sociolingüística etc. (Recalde e Váz-

quez Rozas 2009: 51-53, Romero *et al.* 2017: 123-128). Desenvolveuse así a lingüística de corpus (O'Keeffe e McCarthy 2010), dentro desta, nos últimos 20 anos de maneira particular os córpora de lingua falada (*spoken corpora*) (Raso e Mello 2014) e os córpora de fala (*speech corpora*) (Harrington 2010), estes con aplicacións máis restrinxidas (análise fonética e tecnoloxías da fala, fundamentalmente).

Neste contexto é no que se crea o Corpus Oral Informatizado da Lingua Galega (CORILGA), coa finalidade de recompilar unha mostra representativa da lingua oral actual e poñela á disposición do público en aberto a través da rede. Pretende recoller todo tipo de textos orais, dos que se ofrecen as gravacións de voz e as súas transcricións aliñadas a distintos niveis: a nivel da secuencia fónica e a nivel da palabra no caso da transcrición ortográfica, e a nivel de segmento no caso da transcrición fonética. Actualmente conta con anotacións de tipo morfosintáctico, pero contempla a posibilidade de introducir novos niveis de etiquetado no futuro (pragmático, prosódico, léxico etc.).

O CORILGA presentouse por primeira vez en 2011, daquela aínda como idea de proxecto, no simposio organizado polo ILG, *Textos, palabras e voces: córpora e ferramentas para a investigación lingüística*. Naquel encontro, que conmemoraba o 40 aniversario do ILG, analizouse o traballo desenvolto polo Instituto neses corenta anos, e tamén se discutiron cales debían ser os eixes sobre os que debía xirar a investigación da institución nos anos seguintes. Entre estes eixes, destacouse a necesidade de impulsar os traballos no ámbito da variación e cambio lingüístico e, de maneira especial, os orientados ao estudo da lingua falada actual. Neste sentido, un corpus deste tipo resultaba ser unha peza clave, pois podía fornecer material fundamental que permitise levar a cabo estudos de cambio e contacto lingüísticos, así como investigacións no campo da variación diastrática e diafásica, entre outros.

No tempo transcorrido desde aquel momento, o financiamento doutros proxectos relacionados e, fundamentalmente, a posta en marcha da rede Tecnoloxía e Análise de Datos Lingüísticos (TecAnDaLi), coordinada polo Instituto da Lingua Galega, permitiu desenvolver o corpus e realizar unha interface de busca. No ámbito desta rede, a colaboración do Grupo de Tecnoloxías Multimedia (GTM) da Universidade de Vigo deulle un impulso importantísimo, coa incorporación de tecnoloxías da fala (*vid.* sección 4). O equipo de investigación contou tamén coa colaboración de persoal en formación, persoal informático e outros investigadores que enriqueceron de maneira importante este proxecto. Ademais, moitos estudantes e investigadores doaron o seu material para este corpus ao longo destes anos. Por outro lado, neste tempo xa se realizaron diferentes investigacións cos materiais do

corpus, malia non estar aínda dispoñible en aberto, e foron realizadas presentacións en diferentes congresos.

Neste capítulo expoñeranse os obxectivos do CORILGA, describirase a súa estrutura e o seu contido, así como as ferramentas informáticas e tecnolóxicas que incorpora, e daranse algúns exemplos para a súa explotación.

2. Antecedentes e xustificación do proxecto

Nos corenta anos de historia do ILG pódense consignar moitas achegas valiosas por parte dos investigadores que o conformaron e conforman a día de hoxe, pero sen dúbida foi nos seus inicios, especialmente na súa primeira década de existencia, na que se fixo unha contribución esencial á recollida de datos do galego oral. Son moi numerosos naquela época os traballos académicos en que se gravaron textos de diferentes variedades do territorio de fala galega (Regueira, 2008). Foi tamén nesa década (en concreto entre 1974 e 1977) na que se realizaron as enquisas do Atlas Lingüístico Galego (ALGa) en 167 puntos do dominio lingüístico galego. Nesa recolla sistemática de datos, levada a cabo por Rosario Álvarez, Francisco Fernández Rei e Manuel González González, xunto cos cuestionarios en que requirían informacións léxicas, fonéticas e gramaticais aos informantes, tamén realizaron rexistros sonoros de entrevistas semidirixidas (arredor de 30 horas), que adoitan ser textos monologados de contido etnográfico (Fernández Rei 2008). Pero a contribución do Instituto non só se circunscribe ao material oral recollido para a confección do atlas, senón que tamén foron moitas as teses de licenciatura e de doutoramento que describían unha determinada variedade dialectal para as que se compilaron datos e mostras da lingua falada. Eran as denominadas «teses de falas», que se realizaron ao longo dos anos setenta e oitenta e que constitúen tamén o principal antecedente dos estudos sobre o cambio lingüístico en galego. Os investigadores naquel momento eran conscientes de que a sociedade estaba a sufrir unha importante transformación e que isto inevitablemente estaba a ter importantes consecuencias na lingua. De aí naceu o interese por intentar recoller todo o patrimonio lingüístico que se estaba «perdendo». Por esa razón, eses traballos centráronse principalmente nas falas rurais e houbo que agardar ata 1998 para que aparecese a tese de Francisco Dubert, a primeira, e única polo momento, realizada sobre unha fala urbana e os procesos de variación e cambio lingüístico que nela se estaban a producir (posteriormente publicada en Dubert 1999).

No momento en que se decidiu poñer en marcha o CORILGA existían córpora de fala que se usaran para o estudo da fonética e da entoación e que

deran lugar a un número importante de traballos tanto nos campos citados como no ámbito das tecnoloxías da fala (*vid*. Fernández Rei e Regueira 2017 para unha revisión destes traballos). Trátase de córpora de fala (*speech corpora*), gravados en condicións moi controladas, pensados para facer análises acústicas e deseñados *ad hoc* para recoller mostras do fenómeno que se estuda. Fronte a este tipo de córpora, buscábase a constitución dun corpus de lingua oral (*spoken corpus*), é dicir, un corpus en que o obxectivo non era a análise de características segmentais ou suprasegmentais, senón reunir mostras reais que nos permitisen realizar análises lingüísticas en distintos niveis (*vid*. Romero *et al*. 2017: 127 para a distinción entre córpora de fala e córpora de lingua oral).

Naquela altura xa contabamos con algún corpus de lingua escrita, que recollía unha pequena parte oral:

a) O CORGA (Corpus de Referencia do Galego Actual) é un corpus composto por distintos tipos de texto do período comprendido desde 1975 ata a actualidade. É un corpus escrito de 41 millóns de palabras aproximadamente, mais a partir da súa versión 3.0 conta cun pequeno subcorpus de 50 transcricións ortográficas de programas de radio dos anos 90, en que se ofrece un aliñamento da transcrición coa voz. Os autores pretenden ir incluíndo máis transcricións ortográficas de programas radiofónicos (*vid*. Domínguez Noya *et al*. neste volume).
b) En canto ao TILG (Tesouro Informatizado da Lingua Galega) é un corpus histórico de textos galegos da Idade Moderna e Contemporánea. Na súa versión actual, 4.1 (2018), compila máis de 3000 documentos producidos entre 1612 e 2013 e contén arredor de 31 millóns de palabras gráficas, lematizadas e anotadas con etiquetas morfosintácticas. Deses 31 millóns, case 1,5 millóns de palabras corresponden a transcricións de lingua oral.

Tamén dispoñiamos naquel momento doutro corpus constituído exclusivamente por lingua oral: o AGO (Arquivo do Galego Oral), que contén sobre 2000 horas de gravacións, con voces duns 7000 informantes correspondentes á práctica totalidade dos 315 concellos de Galicia e a boa parte do territorio de fala galega de Asturias, León e Zamora, xunto con mostras das falas dos «tres lugaris» da Serra de Gata, en Cáceres. O AGO consta de cinco subarquivos: tres deles son etnotextos con mostras de galego esencialmente popular (gravacións do ALGa dos anos 70 e textos desde a década de 1980 á actualidade), o cuarto é o subarquivo do Cancioneiro Popular Galego de Schubarth

e Santamarina, e o quinto son textos galegos en rexistro culto (na súa maioría conferencias, pero tamén algún programa radiofónico e algunha serie de televisión) (Fernández Rei 2008).

Naqueles anos tamén se estaba poñendo en marcha un corpus do español de Galicia, ESLORA (*vid*. capítulo 8 neste volume). Este corpus está integrado por entrevistas semidirixidas e conversas espontáneas rexistradas en Galicia entre 2007 e 2015 como parte do Proyecto para el Estudio Sociolingüístico del Español de Galicia (PRESEGAL), que se desenvolve na Universidade de Santiago de Compostela (Vázquez Rozas 2014).

No ámbito do portugués cabe mencionar Corp-ORAL, un corpus de fala espontánea do portugués europeo, desenvolvido polo ILTEC en Lisboa (Freitas e Santos 2008) entre 2005 e 2008. Está constituído por un total de 50 horas de gravación, das que se transcribiron ortograficamente 30. Así mesmo tamén contaba con notación prosódica co fin de permitir o adestramento dos sistemas de recoñecemento e síntese de voz en portugués, favorecer os estudos de fonética e fonoloxía, así como estudos lexicais e gramaticais, e mesmo pragmáticos e sociolingüísticos. Este proxecto tivo continuidade en Oralphon (2010-2012), que engadía a Corp-Oral nova información, como unha ampliación da etiquetaxe do corpus transcrito e a inclusión dunha transcrición fonética estreita (4 horas)[1].

Por tanto, no momento en que se comezaron a recompilar datos e textos para CORILGA, partiamos da existencia dalgúns córpora de textos escritos (CORGA e TILG), un corpus de lingua oral con material non aliñado co son (AGO) e unha cantidade importante de material oral gravado, principalmente textos orais representativos de todas as variedades dialectais con falantes que respondían na súa maioría ao perfil do informante da dialectoloxía tradicional. Como sinalabamos, tratábase de discurso libre e entrevistas semidirixidas nas que participaba normalmente un único falante. Os xéneros predominantes eran as narracións e as descricións, aínda que tamén se recollían contos, romances ou cancións. Se ben este material non era moi variado no que se refería ao estilo e ao xénero, si que estaba dispoñible para incorporar en boa parte a CORILGA e, ademais, posuía un enorme valor: por un lado, porque eran gravacións que nalgún caso tiñan máis de cincuenta anos de antigüidade e, por outra parte, porque en moitas ocasións estaban acompañadas da súa transcrición (*vid*. sección 3).

Este tipo de materiais, a pesar da súa indubidable importancia, estaba deseñado fundamentalmente para estudar a variación diatópica e documen-

[1] O corpus é consultable a través da interface SPOCK (http://spock.iltec.pt/, consultado o 17.02.2020).

tar os recursos e os repertorios da lingua falada seguindo os criterios da dialectoloxía tradicional. Para estudar a variación diastrática e diafásica facíase necesario, por tanto, reunir materiais doutros tipos de texto, doutros ámbitos sociais e doutros rexistros: conversa, lingua de xente nova, lingua urbana, textos formais, entre outros.

Así, CORILGA ten tamén como finalidade cubrir esas necesidades recollendo sobre todo as variedades que estaban menos representadas. En especial, interesábanos ampliar as mostras de conversa, debido á importancia destes textos para a descrición lingüística:

> La conversación coloquial no sólo es el género en el que emergen de forma constante nuevas estructuras lingüísticas, sino también aquel en el que mejor se detectan los procesos de gramaticalización y lexicalización, y, en definitiva, el espacio donde se origina y se hace visible el cambio lingüístico (Recalde e Vázquez Rozas 2009: 53).

Deste xeito, comezou o desenvolvemento de CORILGA e establecéronse como obxectivos principais:

— Reunir un corpus de gravacións transcritas e aliñadas que permita estudar a oralidade contemporánea.
— Recoller a variación existente no galego oral actual, alén da variación diatópica.
— Fornecer material que facilite a análise do cambio lingüístico en tempo aparente e en tempo real.
— Favorecer a elaboración de estudos interdisciplinares, especialmente lingüísticos (morfoloxía, fonética, dialectoloxía, análise do discurso, pragmática etc.).
— Contribuír á creación e ao desenvolvemento de tecnoloxías da fala.

Como veremos ao longo deste artigo, moitos destes obxectivos xa foron alcanzados parcialmente, tanto no que se refire á súa composición e estrutura (sección 3) coma no que atinxe ás súas utilidades e aplicacións (seccións 4 e 5).

3. Contido e estrutura do corpus

O CORILGA incorpora un número aínda reducido de textos, mais xa cobre un importante rango tanto de variación como de espazo temporal. O corpus está constituído por gravacións de diferentes variedades de lingua, que van desde a oralidade formal (conferencias, debates, intervencións parlamentarias, tex-

tos literarios), ata a oralidade informal (monólogos, entrevistas, conversas), pasando polos medios de comunicación (entrevistas, informativos, dobraxes, entre outros). Na táboa 1 mostramos os tipos de texto e as horas de gravación e transcrición dispoñibles actualmente en CORILGA:

Tipo de texto	Horas de gravación	Horas de transcrición
Oralidade informal	34	28
Oralidade formal	53	18
Medios de comunicación	18	9
TOTAL	105	55

Táboa 1. Cómputo de horas de gravación e de transcrición dispoñibles en CORILGA segundo o tipo de texto

O proxecto conta actualmente con 105 horas de gravación e 55 horas transcritas, aínda que este número verase incrementado significativamente nos próximos meses (máis de 40 horas transcritas, 15 delas con transcrición fonética realizada manualmente). Así mesmo, tamén abrangue un amplo período de tempo, desde 1965 ata a actualidade. Os textos recollidos son representativos da variedade existente, de xeito que hai mostras de rexistro oral formal e informal, variedades estándar e non estándar, así como textos provenientes de falantes de xeracións diferentes e niveis socio-culturais distintos.

Estes textos recompílanse a partir de gravacións propias de CORILGA (desde 2012 ata a actualidade), pero tamén a partir de materiais existentes, que foron doados ao ILG ou cedidos para seren incorporados a CORILGA. Algúns dos máis importantes son:

— O corpus doado ao ILG de Gustav Henningsen: son arredor de 100 horas de gravación, que non contan con transcrición. Son gravacións de entrevistas realizadas en diferentes lugares de Galicia entre 1965-1967. Hai textos en galego e en español.
— Gravacións realizadas por Francisco Dubert en Santiago de Compostela entre 1994 e 1996. Son arredor de 20 horas e contan cunha transcrición fonética. Son entrevistas semidirixidas e monólogos.
— Gravacións de Xosé Luís Regueira realizadas para a súa tese de doutoramento en Vilalba, entre 1983 e 1984. Son aproximadamente 16 horas de gravación de monólogos e entrevistas semidirixidas e contan con transcrición fonética.

— Arquivo do Galego Oral: abranguen textos dos anos 1980 aos 2000 e supoñen máis de 100 h de gravación, das que ata o momento se atopan transcritas só 7 h.
— Manuel Rico (entrevistas RNE-Galicia, anos 1980), cunha duración de 16:30 horas.
— Gravacións procedentes de traballos académicos desde 2012 ata a actualidade. Fundamentalmente conversas e entrevistas semidirixidas. Aproximadamente unhas 12 horas de gravación.
— Outras gravacións de diversa procedencia, froito de doazóns (gravacións realizadas para diferentes finalidades, como recollidas de música tradicional, entre outras). Son de aproveitamento limitado, debido aos problemas de calidade que presentan en bastantes casos.

As gravacións transcríbense e anótanse co programa ELAN, (Brugman e Russel 2004, Drude *et al.* 2012), de xeito que se xeran arquivos de transcrición e etiquetado co formato .eaf (Elan Annotation Format). Estes arquivos de transcrición contan con diferentes liñas de anotación para cada informante, aliñadas coa onda sonora:

— Liña de transcrición ortográfica (segmentada por grupos fónicos)
— Liña de transcrición fonética (AFI) (aliñada segmento a segmento)
— Liña de palabra
— Liña de lema
— Anotación morfosintáctica
— Liña de lingua (galego, español)
— Liña de tema
— Liña de tipo de texto

A liña de transcrición principal de CORILGA é a ortográfica (ORT), que se realizou seguindo os criterios habituais nas transcricións de lingua oral para análise do discurso e da conversa (adaptados de Payrató 2003). Por tanto, respéctase fundamentalmente a normativa ortográfica e, asemade, procúrase presentar unha transcrición fiel ao texto oral. Así mesmo, inténtase contribuír a que a recuperación de información sexa o máis sinxela posible, para o que se evitan as variantes ocasionais que aparecen na lingua falada, pois incrementarían notablemente a variación das formas representadas. *Vid.* Dopazo (2018) para unha descrición detallada dos criterios e convencións utilizados na transcrición ortográfica do corpus.

Polo que se refire á transcrición fonética, utilízase o Alfabeto Fonético Internacional (AFI), aínda que cómpre ter en conta que actualmente non están completamente revisadas todas as transcricións, polo que se poden atopar algunhas realizadas de forma automática, que non representan con exactitude as realizacións dos falantes, de aí que se aconselle escoitar o exemplo para verificar que a pronuncia se corresponde coa transcrición.

Para a xestión do material foi creada unha base de datos en MySQL (Ullman *et al.* 2002) cunha serie de táboas:

— Falantes: contén datos e detalles biográficos que se consideran que poden ser pertinentes para realizar estudos comparativos (nivel de estudos, idade, lugar de residencia e nacemento, lingua materna etc.).
— Gravacións: almacena información sobre a data de gravación, o tema, o lugar en que se levou a cabo a gravación etc.
— Temas: contén unha lista de posibles temas (música, vacacións, matrimonio etc.).
— Tipos: contén unha lista de posibles tipos de gravacións (formal, informal, entrevista, monólogo etc.).
— Usuarios: recolle o nome de usuario e o contrasinal das persoas autorizadas para entrar no sistema e engadir ou eliminar gravacións.

A base de datos úsase para almacenar e xestionar información, permitindo buscas rápidas e sinxelas baseadas nos distintos criterios, como tipo de gravación, ano de gravación ou idade do falante, entre outros. Tamén permite conectar datos de distintas táboas de xeito que se poden executar buscas complexas. Así, por medio do filtrado das buscas, poden compararse resultados en textos segundo os diversos tipos e graos de formalidade, para o estudo da variación, así como segundo as diferentes características dos informantes. Utilizando os filtros de data da gravación e do tramo de idade dos informantes, o CORILGA permite tamén o estudo do cambio lingüístico tanto en tempo aparente como en tempo real.

A interface para executar as buscas consta de dous bloques principais: un buscador e unha zona de administración, á cal só teñen acceso as persoas con privilexios. O buscador componse de tres partes principais: a selección dos arquivos nos que se vai realizar a procura, a definición dos patróns de busca e a visualización dos resultados.

Selección de ficheiros (figura 1). Ofrécese a posibilidade de realizar procuras dentro de todos os arquivos existentes na base de datos ou ben facelo só dentro dos que cumpran determinados requisitos: ano ou período

en que se realizou a gravación, o tipo de texto, o hábitat (urbano, rural ou semiurbano) e o tema (por exemplo, historias de vida, lingua e comentarios sociolingüísticos, literatura, traballos do agro, oficios, política, relixión, festas etc.). Tamén se pode aplicar un filtrado polas características dos falantes: idade, sexo, nivel de estudos, lingua inicial, lingua da gravación, lugar de nacemento ou de residencia. Desta maneira poden filtrarse falantes por xeracións en diferentes momentos temporais, grupos de sexo, procedencia, e en diferentes situacións comunicativas e diferentes tipos de lingua.

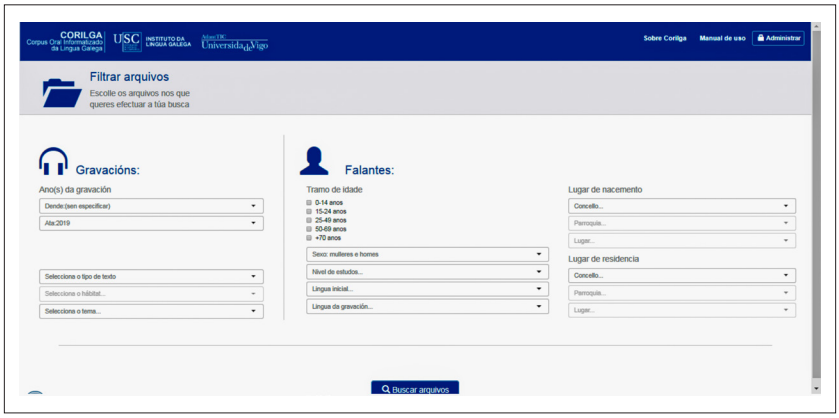


FIGURA 1. Menú de selección de gravacións e falantes

Unha vez feita a escolla dos arquivos e falantes nos que queremos levar a cabo a busca, podemos **definir os patróns de busca** (figura 2), co fin de executar as procuras nas diferentes liñas (tiers). Para definir as buscas que se van realizar dentro das liñas dos distintos ficheiros que foron seleccionados no paso anterior, a aplicación ofrece a posibilidade de facer a busca dun patrón nunha liña, pero tamén permite buscar en varias liñas simultaneamente.

A busca executada devolve unha **táboa de resultados**, onde se listan as coincidencias atopadas, que se poden escoitar, pausar, expandir, borrar e descargar en formato Excel (transcricións e anotacións), Praat (o fragmento de audio), ou ELAN (audio, transcricións e anotacións aliñadas), segundo a finalidade que se pretenda (figura 3). Tamén se ofrece a posibilidade de levar a cabo unha descarga múltiple, que permite seleccionar diferentes coincidencias mediante os cadros selectores que se atopan ao inicio de cada liña de coincidencia e descargalas todas xuntas nun mesmo arquivo .zip (para ELAN

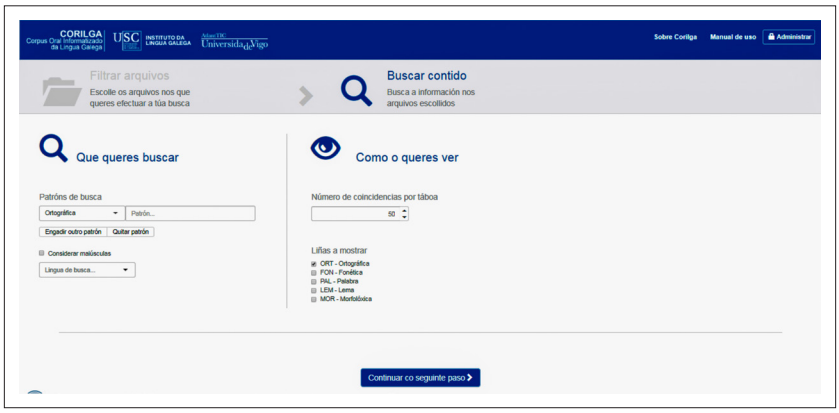


FIGURA 2. Menú de busca de variables lingüísticas

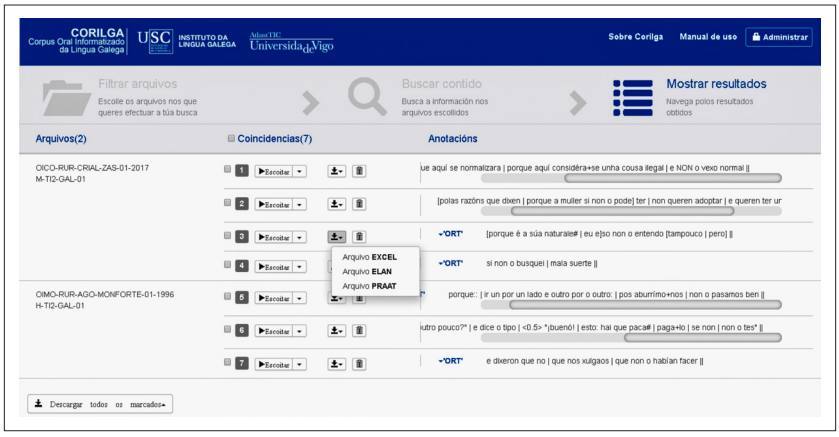


FIGURA 3. Imaxe dunha táboa de resultados da busca dunha secuencia. Móstranse as coincidencias, as anotacións e os formatos de descarga (EXCEL, ELAN e PRAAT)

ou Praat) ou un Excel que contén todas as coincidencias seleccionadas, con cada resultado coincidente identificado polos códigos de ficheiro e falante.

4. FERRAMENTAS INCORPORADAS E DESENVOLVIDAS

Este proxecto, como se dixo anteriormente, desenvólvese en colaboración co grupo de investigación Tecnoloxías Multimedia (GTM) da Universidade de Vigo. Este grupo é pioneiro no desenvolvemento de tecnoloxías da fala, e especificamente no tocante ás tecnoloxías aplicadas á lingua galega. Por tanto, esta colaboración foi de grande importancia para o desenvolvemento

do CORILGA, de igual maneira que este corpus tamén contribuíu a mellorar as tecnoloxías de recoñecemento automático da fala en galego. Información sobre o proceso de deseño e desenvolvemento tecnolóxicos pode encontrarse en García-Mateo *et al.* (2014) e Seara *et al.* (2016).

O programa dispón dunha serie de ferramentas que permiten o aliñamento texto-voz (ao nivel do segmento, da palabra e do grupo fónico) e unha ferramenta de recoñecemento e transcrición automática. O sistema de recoñecemento automático da fala (ASR) está baseado no Kaldi Speech Recognition Toolkit (Povey *et al.* 2011). Estes recursos están complementados cun lematizador e cun etiquetador morfolóxico automático; neste caso, utilizamos o etiquetador Freeling (Padró 2011; Padró e Stanislovsky 2012), na versión adaptada para a lingua galega. Así mesmo, o programa fornece unha transcrición fonética automática, que presenta importantes limitacións por estar realizada a partir da anotación ortográfica (e por tanto non ten en conta características fonéticas relevantes, como o grao de abertura das vogais medias), mais que serve como unha transcrición de base que permite aforrar tempo no lento proceso de transcrición fonética.

O resultado obtido de todas estas ferramentas é un arquivo eaf do programa Elan, con anotacións multinivel sincronizadas co arquivo de audio ou audio e vídeo (*vid.* sección 3), tal e como mostra a figura 4.

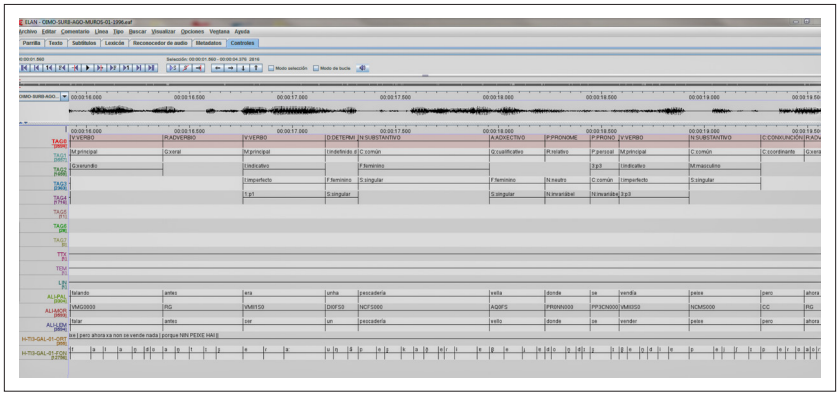


FIGURA 4. Imaxe dun segmento dun arquivo eaf cos distintos niveis de anotación aliñados co arquivo de audio

A ferramenta web está automatizada, de maneira que abonda con subir (*drag and drop*) un arquivo de son con formato wav para que o eaf con todas as liñas transcritas e anotadas se xere automaticamente. Se se conta cunha transcrición ortográfica parcial do texto, pode engadirse nun arquivo eaf, e nese caso

o recoñecedor toma en consideración ese texto como modelo de lingua e pode mellorar notablemente a calidade da transcrición automática (figura 5).

FIGURA 5. Pantalla de recoñecemento automático de fala

No caso de contar xa cunha transcrición ortográfica fiable pode engadirse en formato eaf ou txt, e nese caso o programa aliña a transcrición fornecida co arquivo de audio e completa o resto das liñas de anotación. E de dispoñer tamén dunha transcrición fonética (como ocorre en gravacións que proceden de teses de doutoramento, para as que se realizaron transcricións fonéticas), pode engadirse en formato txt para que o programa a aliñe segmento a segmento (figura 6).

FIGURA 6. Pantalla de aliñamento de transcricións ortográfica e fonética

As tecnoloxías da fala son instrumentos de grande importancia para o funcionamento do CORILGA, como se acaba de ver. Mais, por outra parte, este corpus tamén supón unha contribución ao desenvolvemento destas tecnoloxías aplicadas ao galego. En relación co desenvolvemento deste proxecto, leváronse a cabo traballos de mellora dos modelos de recoñecemento da fala, en parte con textos deste mesmo corpus, aínda que tamén con outros externos (Piñeiro *et al.* 2018). Neste traballo, no que se utilizou o SRI Language Modeling Toolkit (SRILM) sobre a plataforma de recoñecemento automático de fala Kaldi, mostrouse como o adestramento con diferentes modelos de lingua ten un impacto diferente nos resultados obtidos, e neste caso supuxo unha notable mellora tanto nas ratios de erro de palabra (WER, Word error rate) coma nas formas non recoñecidas por estaren fóra do vocabulario (OOV, Out of vocabulary).

O experimento realizouse con tres córpora:

a) Primeiro Corpus: oralidade formal (30 arquivos dunha duración media de 3:50 minutos, cunha duración total de 115 minutos).
b) Segundo Corpus: programas de noticias da Televisión de Galicia (10 arquivos de duración media de 34 minutos, cunha duración total de 340 minutos).
c) Terceiro Corpus: TED Talks (10 arquivos de duración media de 16 minutos, e unha duración total de 163 minutos).

Eses tres córpora foron tratados con diferentes modelos lingüísticos ou con combinacións de modelos:

— DOG (Diario Oficial de Galicia): tamaño medio, lingua moi formal.
— DUVI (Diario da Universidade de Vigo): tamaño pequeno, mais representativo e actual.
— BEPUB: gran cantidade de novelas traducidas con tradutor automático ao galego. Calidade limitada, mais de gran tamaño.
— Wikipedia: representativo e variado, mais non contén material lingüístico informal.
— GEV (Diario dixital Galicia Hoxe, páxina de noticias de Eroski e diario dixital Vieiros): lingua actual, variada e corpus extenso.
— CORGA (Corpus de Referencia do Galego Actual): tamaño medio, corpus representativo.
— Suma 1 (texto de CORGA, GEV e DUVI).

— Suma 2 (Modelos adestrados de CORGA, GEV e DUVI).
— Suma 3 (Modelos adestrados de CORGA e BEPUB).
— Suma 4 (Modelos adestrados de CORGA, GEV, DUVI e BEPUB).

Os resultados do adestramento con estes modelos pode verse na figura 7.

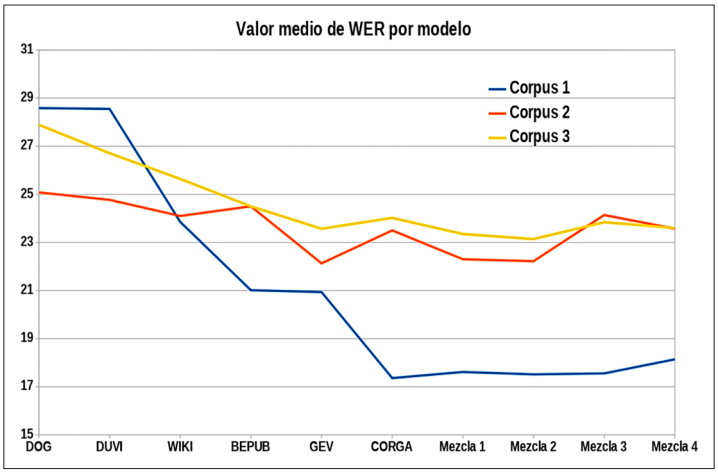


FIGURA 7. Evolución da WER media para os diferentes marcos experimentais (tomado de Piñeiro *et al.* 2018)

Como se pode comprobar, o corpus 1 (oralidade formal, con textos do CORILGA) experimenta unha redución significativa dos erros de identificación léxica, mentres que esa redución é moito menor nos outros dous córpora. Esta diferenza de comportamento é rechamante, e cabe formular diferentes hipóteses sobre as posibles causas (*cfr.* Piñeiro *et al.* 2018: 81). Da análise destes comportamentos e dos adestramentos necesarios de modelos lingüísticos deberán derivarse novas melloras das tecnoloxías de recoñecemento de fala.

Así pois, o CORILGA beneficiase en grande medida das tecnoloxías de fala dispoñibles para o galego e asemade tamén contribúe á súa mellora e ao seu refinamento.

5. Posibilidades de explotación do Corpus

Aínda que o CORILGA non está aínda accesible de maneira pública, algúns traballos xa se están a basear nos materiais lingüísticos deste corpus. Un exemplo é o traballo presentado por Esther Brown e Javier Rivas (2018)

sobre a variación no uso do infinitivo conxugado entre o discurso formal e a lingua coloquial informal. Tamén se utilizou parcialmente este corpus para outros traballos sobre contacto de linguas no discurso político (Regueira 2016; en prensa) ou sobre a variación lingüística na construción de personaxes nun programa de humor na Televisión de Galicia (Varela 2016), entre outros. Actualmente están en marcha traballos sobre estruturas posesivas ou sobre estratexias de cortesía en entrevistas de medios de comunicación, entre algúns outros, que se basean parcialmente nos textos contidos neste corpus. Isto mostra que xa empezou a cumprir o principal obxectivo marcado desde o seu inicio: fornecer un corpus de materiais lingüísticos que posibilite ou facilite a investigación lingüística sobre o galego oral actual, especialmente nos eidos da variación e do cambio lingüístico. Sen dúbida, a súa funcionalidade verase grandemente incrementada coa incorporación dun número moito maior de textos de diferentes tipos, prevista para unha próxima fase de traballo, de maneira que poida ser posto á disposición pública no ano 2020.

Debido á súa estrutura e á información lingüística dispoñible, o CORILGA permite realizar diferentes tipos de traballos nos ámbitos indicados (variación e cambio lingüístico):

a) Variación entre lingua formal e lingua informal, lingua dos medios, lingua da actividade pública (política, cultura...) variación oral / escrito (lectura), entre outras.
b) Cambio en tempo aparente: selección de falantes de diferentes tramos de idade (xeracións) nun mesmo intervalo temporal (p. e. falantes do tramo 25-50 anos / vs / falantes de 50-70 anos, en gravacións na década de 1990).
c) Cambio en tempo real: un mesmo tramo de idade en dous momentos temporais diferentes (p. ex. falantes de 25-50 anos en gravacións dos anos 1960 e en gravacións dos anos 1990).
d) Estudos lonxitudinais: estudo dun mesmo grupo de idade ao longo do tempo.

Levar a cabo estes estudos é posible a través da utilización dos filtros da información sociolingüística que permiten seleccionar só as gravacións ou as persoas que reúnan determinadas condicións, como xa se expuxo na sección 3 (*vid.* figura 1).

Diferentes aspectos da variación lingüística no galego das ultimas décadas poden ser estudados con este corpus. Por poñer só uns poucos exemplos, poden facerse buscas sobre o infinitivo conxugado en diferentes épocas

e diferentes variedades de lingua, combinando p. ex. a secuencia ortográfica «rmos» coa etiqueta «Infinitivo conxugado + p4» (deixando fóra do filtro desta maneira secuencias coincidentes, como o substantivo «termos» ou «enfermos»). Desta forma pode someterse a proba a hipótese de que na lingua actual o infinitivo conxugado é utilizado como un marcador de formalidade do discurso.

Outro aspecto de interese é a perda do clítico /no/ (co seu feminino e os correspondentes plurais), substituído pola forma non marcada /o/, que se detecta en xente nova e nos estilos formais, como en *non o*, pronunciado [nõŋʊ] fronte a [nõnʊ] (análise morfolóxica [NEG non] + [CLIT o], fronte a [NEG no(n)] + [CLIT no]; tamén sucede o mesmo en *quen o*, *tamén o*, *ben o* etc. Estas secuencias están diferenciadas xa na liña ortográfica, así que pode buscarse «non o» fronte a «non-o», e na etiqueta morfolóxica «pronome persoal» (para deixar fóra os posibles casos en que «o» é artigo).

Tamén son aspectos de interese na variación lingüística o uso das formas de subxuntivo *cantara ~ cantase*, o uso das perífrases *estar a facer ~ estar facendo*, a colocación dos clíticos en casos como *para lle facer ~ para facerlle*, o uso das formas de tratamento *usté ~ vostede ~ vosté*, só por poñer algúns exemplos.

Este tipo de traballos ten, fundamentalmente, interese teórico, mais tamén pode ter aplicacións en diferentes terreos, como a tradución ou o ensino/aprendizaxe da lingua (sobre posibles utilizacións na aula de lingua galega como L1 ou como L2 *vid*. Regueira 2017: 294-297).

6. Conclusións

O CORILGA pretende, segundo se defendeu neste capítulo, converterse nunha ferramenta para potenciar os estudos sobre a variación e o cambio lingüístico no galego falado, aínda que tamén serve para estudar outros aspectos relacionados, como o contacto lingüístico co español (e, en menor medida, co portugués), entre outros varios. A idea que o sustenta é fornecer un corpus lingüístico que poida servir de base para traballos de investigación sobre algún destes aspectos. Ata o momento, as persoas que pretendían iniciar algunha investigación nestas liñas tiñan que construír o seu propio corpus, e nesa tarefa consumían unha gran cantidade de tempo e esforzo. Evitar ou aliviar esa situación é a principal finalidade do corpus que estamos a describir, e que aspira a converterse, por tanto, nunha potente ferramenta para o estudo da variación sociolingüística na lingua das últimas décadas, así coma do cambio lingüístico e do contacto de linguas. Ademais de ter outras apli-

cacións potenciais (para o ensino da lingua, por exemplo), este corpus quere contribuír ao desenvolvemento e á mellora das tecnoloxías da fala, especialmente o recoñecemento de fala e a transcrición automática.

Somos conscientes de que o carácter aberto deste corpus e a ambición de recoller a variación social e estilística do galego actual constitúen o que se podería considerar a súa principal debilidade: un corpus demasiado ambicioso que require unha gran cantidade de tempo e de esforzo para reunir unha cantidade relevante de texto transcrito e anotado. Resulta obvio que o financiamento dun recurso deste tipo é moi dificultoso, na situación actual da investigación universitaria, mais contamos coa axuda das tecnoloxías da fala e doutras ferramentas de anotación para reducir significativamente a carga de traballo. Na primeira fase, que esperamos ter completada a principios do ano 2020, a cantidade de fala non será aínda moi grande, mais cremos que o corpus será xa de utilidade para a investigación.

Outra limitación deste corpus é que só permite recuperar fragmentos de fala de curta duración, e non textos nin conversas completas. Por tanto, a súa utilidade é limitada para os estudos de discurso ou a análise da conversa. Non obstante, para este tipo de traballos o equipo responsable do corpus permite acceder aos textos completos baixo demanda, polo que esta pexa pode ser superada.

Este carácter aberto fai que teña tamén outra potencialidade non menor, como sería que nun futuro poida combinarse con outros córpora do español de Galicia, como o corpus ESLORA, de maneira que poida ser estudada a variación lingüística en Galicia, superando a separación entre galego e español en compartimentos estancos. Unha comparación co portugués, con córpora orais como o CORP-ORAL, tamén é posible, en canto que tamén se traballa con transcricións aliñadas cos mesmos programas informáticos.

Agradecementos

A elaboración do CORILGA desenvolveuse, na súa maior parte, dentro da Rede Tecnoloxías e Análise dos Datos Lingüísticos (TecAnDaLi, http://ilg.usc.es/tecandali/), que contou con financiamento parcial do Fondo Europeo de Desenvolvemento Rexional (FEDER). Unha parte deste corpus desenvolveuse no marco dos proxectos de investigación «Cambio lingüístico no galego actual», e «Contacto e cambio lingüístico en galego», financiados polo Ministerio de Economía e Competitividade (FFI2012-33845 e FFI2015-65208-P, respectivamente).

Recursos electrónicos

CORGA: Corpus de Referencia do Galego Actual http://corpus.cirp.gal/corga/ [consultado o 16.05.2019].

ELAN: The Language Archive, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands https://tla.mpi.nl/tools/tla-tools/elan/ [consultado o 16.05.2019].

FreeLing: http://nlp.lsi.upc.edu/freeling/index.php/node/1 [consultado o 16.05.2019].

FreeLing Galego: Análise lingüística automática do galego http://sli.uvigo.gal/lingua/ [consultado o 16.05.2019].

Praat: Boersma, Paul & David Weenink (2019). Praat: doing phonetics by computer http://www.praat.org/ [consultado o 16.05.2019].

SPOCK: The Spoken Corpus Klient http://spock.iltec.pt/ [consultado o 16.05.2019].

SRILM: SRI Language Modeling Toolkit. http://www.speech.sri.com/projects/srilm/ [consultado o 15.05.2019].

TECANDALI: Rede Tecnoloxías e Análise dos Datos Lingüísticos http://ilg.usc.es/tecandali/ [consultado o 15.05.2019].

TILG: Tesouro Informatizado da Lingua Galega http://ilg.usc.gal/TILG/ [consultado o 16.05.2019].

Referencias bibliográficas

Brown, Esther & Javier Rivas (2018): «A quantitative variationist analysis of Galician inflected infinitives», Relatorio presentado no *III North American Symposium of Galician Studies: Galician Studies Moving West*. Denver: Regis University / Metropolitan State University of Denver / Colorado College, 18-20 de outubro de 2018.

Brugman, Hennie & Albert Russel (2004): «Annotating Multimedia / Multi-modal resources with ELAN», in *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*. Lisboa: ELRA.

Dopazo Entenza, José Manuel (2018): «Construyendo un corpus oral para el gallego: el proyecto CORILGA», *e-Scripta Romanica* 5, pp. 28-38.

Dubert García, Francisco (1999): *Aspectos do galego de Santiago de Compostela*. Santiago de Compostela: Universidade de Santiago de Compostela.

Drude, Sebastian, Daan Broeder, Paul Trilsbeek & Peter Wittenburg (2012): «The Language Archive: A new hub for language resources», in Nicoletta Calzolari (ed.): *Proceedings of LREC 2012: 8th International Conference*

on Language Resources and Evaluation. Istambul: European Language Resources Association (ELRA), pp. 3264-3267.

Fernández Rei, Elisa & Xosé Luís Regueira (2017): «Situando o galego no terreo da investigación lingüística: os traballos de fonética e fonoloxía», *LaborHistórico* 3/1, pp. 93-110. https://doi.org/10.24206/lh.v3i1.17109

Fernández Rei, Francisco (2008): «O ALGa e o Arquivo do Galego Oral: a imaxe do galego», in Elisa Fernández Rei & Xosé Luís Regueira Fernández (eds.): *Perspectivas sobre a oralidade*. Santiago de Compostela: Consello da Cultura Galega / Instituto da Lingua Galega, pp. 35-74.

Freitas, Tiago & Fabíola Santos (2008): «CORP-ORAL: a spontaneous European Portuguese speech resource», in *Proceedings. 8th Language Resources and Evaluation Conference (LREC 2008)*. Marrakesh: LREC. http://www.lrec-conf.org/proceedings/lrec2008/; http://www.iltec.pt/pdf/co/co4.pdf

García-Mateo, Carmen, Antonio Cardenal, Xosé Luís Regueira, Elisa Fernández Rei, Marta Martinez, Roberto Seara, Rocío Varela & Noemi Basanta (2014): «CORILGA: a Galician multilevel annotated speech corpus for linguistic analysis», in *Proceedings. 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, pp. 2653-2657. http://www.lrec-conf.org/proceedings/lrec2014/pdf/739_Paper.pdf

Gumperz, John J. (1982): *Discourse strategies*. Cambridge: Cambridge University Press.

Gumperz, John J. & Dell H. Hymes (eds.) (1964): *The ethnography of communication*. Menasha, Wisc.: American Anthropological Association.

Gumperz, John J. & Dell H. Hymes (1972): *Directions in sociolinguistics: The ethnography of communication*. New York: Basil Blackwell.

Harrington, Jonathan (2010): *Phonetic analysis of speech corpora*. Chichester: Wiley-Blackwell.

Labov, William (1966): *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Labov, William (1972): *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

O'Keeffe, Anne & Michael McCarthy (eds.) (2012): *The Routledge handbook of corpus linguistics*. London: Routledge.

Padró, Lluis (2011): «Analizadores multilingües en FreeLing», *Linguamatica* 3/2, pp. 13-20.

Padró, Lluis & Evgeny Stanilovsky (2012): «FreeLing 3.0: Towards wider multilinguality», in *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. Istanbul: ELRA. http://nlp.lsi.upc.edu/publications/papers/padro12.pdf

Payrató, Lluís (2003): *Pragmática, discurs i llengua oral. Introducció a l'anàlisi funcional de textos*. Barcelona: Universitat Oberta de Catalunya.

Piñeiro Martín, Andrés, Carmen García-Mateo, Laura Docío-Fernández & Xosé Luís Regueira (2018): «Estudio sobre el impacto del corpus de entrenamiento del modelo de lenguaje en las prestaciones de un reconocedor de habla», *Procesamiento del Lenguaje Natural* 61, pp. 75-82.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer & Karel Veselý (2011): «The Kaldi speech recognition toolkit», in *2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2011)*. Big Island, Hawaii. https://www.danielpovey.com/files/2011_asru_kaldi.pdf

Raso, Tommaso & Heliana Mello (eds.) (2014): *Spoken corpora and linguistic studies*. Amsterdam: John Benjamins.

Recalde, Montserrat & Victoria Vázquez Rozas (2009): «Problemas metodológicos en la formación de corpus orales», in Pascual Cantos Gómez & Aquilino Sánchez Pérez (eds): *A survey of corpus-based research*. Murcia: Asociación Española de Lingüística del Corpus, pp. 51-64. https://www.um.es/lacell/aelinco/contenido/pdf/4.pdf

Regueira, Xosé Luís (2008): «Os estudos de dialectoloxía galega desde 1967 á actualidade», in Ester Corral Díaz, Lydia Fontoira & Eduardo Moscoso Mato (eds.): *A mi dizen quantos amigos ei: homenaxe ao profesor Xosé Luís Couceiro*. Santiago de Compostela: Universidade de Santiago de Compostela, pp. 573-584.

Regueira, Xosé Luís (2016): «La lengua de la esfera pública en situación de minorización: español y portugués como lenguas de contacto en el lenguaje político gallego», in Dolors Poch Olivé (ed.): *El español en contacto con las otras lenguas peninsulares*. Madrid / Frankfurt am Main: Iberoamericana / Vervuert, pp. 39-59.

Regueira, Xosé Luís (2017): «El aprendizaje de la pronunciación y de la lengua oral: dos herramientas para el aula de lengua gallega», in María José Domínguez Vázquez & María Teresa Sanmarco Bande (eds.): *Lexicografía y didáctica*: *diccionarios y otros recursos lexicográficos en el aula*. Frankfurt am Main: Peter Lang, pp. 283-303.

Regueira, Xosé Luís (en prensa): «Portuguese as a contact language in Galicia: convergence, divergence, ideology and identity», in Miriam Bouzouita, Renata Enghels & Clara Vanderschueren (eds.): *Convergence and divergence in Ibero-Romance: Case studies from the Ibero-Romance world*. Berlin: De Gruyter Mouton.

Romero, Asier, Irati de Pablo, Aintzane Etxebarria & Ainara Romero (2017): «Teorización sobre la construcción de corpus orales en la adquisición del lenguaje: una propuesta metodológica para el procesamiento de lenguas con soporte tecnológico», *Analecta Malacitana Electrónica* 42, pp. 123-155. http://www.anmal.uma.es/AnMal42/Corpus_orales.pdf

Sacks, Harvey, Emanuel A. Schegloff & Gail Jefferson (1974): «A simplest systematics for the organization of turn-taking for conversation», *Language* 50, pp. 696-735.

Seara, Roberto, Marta Martínez, Rocío Varela, Carmen García-Mateo, Elisa Fernández Rei & Xosé Luís Regueira (2016): «Enhanced CORILGA: introducing the automatic phonetic alignment tool for continuous speech», in *Proceedings. 10th Language Resources and Evaluation Conference (LREC 2016)*. Portorož, Slovenia, pp. 2893-3898. http://www.lrec-conf.org/proceedings/lrec2016/pdf/1074_Paper.pdf

Ullman, Jeffrey D., Hector Garcia-Molina & J. Windom (2002): *MySQL: Database systems: the complete book*. Upper Saddle River, NJ.: Prentice-Hall.

Varela, Sonia (2016): «Contacto de lenguas y caracterización de personajes en la Televisión de Galicia (TVG). Análisis de un sketch del programa de humor Land Róber», in Dolors Poch Olivé (ed.): *El español en contacto con las otras lenguas peninsulares*. Madrid / Frankfurt am Main: Iberoamericana / Vervuert, pp. 61-80. https://doi.org/10.31819/9783954878635-004

Vázquez Rozas, Victoria (2014): «ESLORA: Diseño, codificación y explotación de un corpus oral de español de Galicia», in *II Workshop de Procesamiento Automatizado de Texto y Corpus (WOPATEC-2014)*. Pontificia Universidad Católica de Valparaíso, Viña del Mar, 13-14 de noviembre de 2014.

PROBLEMAS AFRONTADOS EN LA ETIQUETACIÓN MORFOSINTÁCTICA DEL CORPUS ESLORA[1]

Problems encountered in the morphosyntactic tagging of the ESLORA corpus

Eva M.ª Domínguez Noya, Raquel Rivas Cabanelas,
M.ª Paula Santalla del Río, Rebeca Villapol Baltar
Universidade de Santiago de Compostela

Resumen

El corpus para el estudio del español oral ESLORA ha sido lematizado y etiquetado morfosintácticamente con el etiquetador XIADA, originalmente desarrollado para el análisis del gallego, pero adaptado para trabajar con español en el marco del proyecto. En este trabajo se examinan algunas cuestiones relacionadas con la adaptación de este etiquetador. Una parte fundamental del proceso llevado a cabo ha consistido en la elaboración del corpus de entrenamiento a partir del cual el etiquetador infiere qué análisis corresponde a cada unidad gramatical en cada caso concreto. En este sentido, el aspecto que entraña más dificultad y que se aborda en este trabajo es la determinación de las etiquetas que deben corresponder a dos tipos de ítems: a) los propios del corpus oral registrado, cuya codificación no está tan inequívocamente sustentada por soluciones descritas en trabajos previos y b) los ítems compartidos por lengua oral y lengua escrita, que poseen usos muy distintos en cada una de ellas, y cuya descripción está sobre todo respaldada por la información gramatical prevista para la lengua escrita.

[1] El proyecto de investigación ESLORA+ (El corpus ESLORA de español oral: enriquecimiento, análisis lingüístico y extracción de recursos, ref. PFFI2017-86379-P) está financiado por la Agencia Estatal de Investigación (AEI) y por el Fondo Europeo de Desarrollo Regional (FEDER). El equipo del proyecto forma parte del grupo de investigación Gramática del español de la Universidade de Santiago de Compostela, beneficiario de una ayuda para «Consolidación e estruturación de Grupos con Potencial de Crecemento 2017» de la Consellería de Cultura, Educación e Ordenación Universitaria de la Xunta de Galicia (ref. ED431B 2017/39).

Palabras clave: corpus oral, anotación morfosintáctica, etiquetario, corpus de entrenamiento

Abstract

The corpus for the study of spoken Spanish ESLORA has been lemmatized and morphosyntactically tagged by means of the POS-tagger XIADA, originally designed for Galician and adapted to Spanish in the context of the ESLORA Project. This study examines some of the problems related to the adaptation of this POS-tagger. A crucial part of the adaptation process involves the development of a training corpus in order to help the POS-tagger infer which analysis corresponds to which word in each specific case. What is most problematic in this process is the correct tagging of two types of linguistic items: a) those that are specific of the oral corpus data and therefore are often not fully, or even not at all, accounted for in previous descriptions, and b) those that occur in speech and in writing but are used very differently, with the existing descriptions being primarily based on written language.

Keywords: spoken corpus, corpus annotation, POS tagging, tagset, training corpus

I. Introducción

El corpus ESLORA es un corpus de lengua oral de español de Galicia constituido por 647 758 palabras de entrevistas semidirigidas y conversaciones que fueron grabadas entre 2007 y 2015. El corpus ha sido transcrito ortográficamente y la transcripción se ha alineado con las grabaciones de audio (una descripción más detallada de distintos aspectos del corpus se recoge en Barcala *et al.* 2018). ESLORA también ha sido enriquecido con anotación lingüística de tipo morfosintáctico. De este enriquecimiento es de lo que trata este trabajo.

Como señalan Graña (2000) y, siguiendo su senda, Domínguez (2013, 2016) y García (2014), en la anotación morfosintáctica de corpus con un etiquetador morfosintáctico probabilístico hay que tener en cuenta el tamaño y la calidad del corpus de entrenamiento (a mayor tamaño y mejor calidad, mejor aprendizaje del modelo), el conjunto de etiquetas y el diccionario. En el apartado 2 nos referiremos a la anotación empleada en ESLORA para estos tres elementos. Lo que queremos destacar es la necesidad, reconocida por todos estos autores, de que haya congruencia entre el corpus de entrenamiento y el corpus anotado desde el punto de vista del registro y el género textual. No solo importa que el corpus de entrenamiento sea grande y esté cuidadosa y consistentemente anotado, sino que es importante también que sea del mismo tipo que el texto que se vaya a anotar. Este hecho lo encon-

tramos también refrendado ya en Manning (2011), en donde el autor se pregunta cómo aumentar la precisión de los etiquetadores más allá del que es el máximo alcanzado en el estado del arte en el momento, máximo al que se llega, señala Manning repetidas veces, cuando hay congruencia en temática, época y estilo entre los datos presentes en el corpus de entrenamiento y aquellos sobre los que se va a operar.

Sin embargo, a lo largo del todavía breve período de tiempo en que se han desarrollado corpus orales en español no ha sido siempre posible etiquetarlos sirviéndose de otros corpus de lengua oral. Solo muy recientemente se ha empezado a disponer de anotación morfosintáctica para los corpus de lengua hablada. Así, el corpus Val.Es.Co[2] incorpora en su aplicación de consulta el uso de etiquetas morfosintácticas, si bien esta opción se califica en el propio corpus como experimental. Val.Es.Co ha sido etiquetado con el etiquetador FreeLing y a su documentación se remite para el manejo en la interfaz de consulta de todo lo relacionado con las etiquetas.

También el *Corpus oral y sonoro del español rural* (COSER)[3] incorpora la modalidad de consulta avanzada que permite hacer búsquedas por etiquetas morfosintácticas, aunque solo desde 2017. No consta en la interfaz en línea de consulta de COSER información acerca de la herramienta o las etiquetas utilizadas en su anotación, pero la facilidad de búsqueda integrada en ella permite seleccionar clases, subclases de palabras y valores para categorías morfosintácticas acordes con las recomendaciones y el uso generales. En Carlota de Benito *et al.* (2016) se explica, en cambio, que la anotación morfosintáctica de COSER se llevó a cabo con el etiquetador FreeLing, adaptando el tokenizador, el módulo de sufijos y, sobre todo, el diccionario de la herramienta para la introducción del léxico de la vida rural que aparece en las transcripciones, así como las realizaciones no estándares así recogidas en la transcripción y ya desde la misma asociadas con sus correspondientes formas estándares. No se modifica, sin embargo, el módulo estadístico de FreeLing y, por lo tanto, no se etiqueta recurriendo a un corpus de entrenamiento específico para el registro oral, sino sobre la base del registro escrito.

Es distinto en este sentido el caso del corpus C-ORAL-ROM[4], que ha sido anotado morfosintácticamente para un conjunto de 129 etiquetas por medio del analizador GRAMPAL, cuya estrategia de desambiguación combina reglas y estadística. GRAMPAL fue diseñado originalmente para analizar textos

[2] http://www.valesco.es/?q=palabras
[3] http://www.corpusrural.es/#
[4] http://cartago.lllf.uam.es/coralromdemo/concor.cgi?m=2

escritos, pero fue progresivamente adecuado para su aplicación a lengua oral en la etiquetación del corpus C-ORAL-ROM. Tras este proceso, los autores (Guirao Miras & Moreno Sandoval 2006) señalan, por ejemplo, que la distribución entre palabras no ambiguas y ambiguas en un corpus escrito es de 65 % a 35 %, respectivamente, mientras que la relación de ambigüedad en un corpus oral está casi al 50 %, y concluyen que, de acuerdo con su experiencia, desde la identificación de unidades de referencia, pasando por el conjunto de etiquetas y el diccionario, hasta las reglas o el corpus de entrenamiento, si los hay, lo que se usa para etiquetar lengua oral debe estar orientado a ella para la obtención de resultados que equilibren razonablemente la calidad y la precisión de la anotación.

El *Corpus del Español del Siglo XXI* (CORPES)[5] también incluye la etiquetación morfosintáctica tanto en sus materiales orales, como en los escritos, con un grado de detalle considerable: 330 etiquetas que han sido aplicadas al corpus por una herramienta estadística con un resultado de acierto en los niveles habituales (97-98 %) (Rojo 2016). No tenemos noticia, sin embargo, de que los recursos utilizados por la herramienta se hayan adaptado para la etiquetación del CORPES escrito por un lado y del CORPES oral por otro, y de ahí que búsquedas centradas en algunos de los ítems tratados en esta exposición ofrezcan resultados que en la etiquetación del CORPES oral podrían ser, creemos, refinados (*vid.* apartado 4.2). También Mendes *et al.* (2004) etiquetan el corpus C-ORAL-ROM para portugués entrenando el etiquetador de Brill con un corpus de lengua escrita. Aunque se muestran satisfechos con los resultados, dada la minimización de costes que para ellos supone esta aproximación, concluyen que una futura mejora de tales resultados deberá ligarse a un entrenamiento del etiquetador con el corpus oral manualmente revisado. La contrapartida, por fin, para el portugués brasileño del corpus anterior, el C-ORAL-BRASIL, se ha etiquetado con el etiquetador PALAVRAS, con resultados muy satisfactorios a pesar de ser considerado como un etiquetador para lengua escrita. Las circunstancias de esta reconversión son, sin embargo, muy particulares por comparación con las de otros anotadores mencionados, pues PALAVRAS es en primera instancia un etiquetador basado en reglas que trabajan con anotaciones de naturaleza cada vez más elaborada, y aunque no se adaptaron las reglas, sí se adaptaron o introdujeron módulos que tienen muchas repercusiones en su funcionamiento (Bick *et al.* 2012).

A la vista de estos antecedentes, cuando se emprendió la anotación del corpus ESLORA se decidió que los recursos utilizados para etiquetarlo

[5] http://web.frl.es/CORPES/view/inicioExterno.view

habrían de ser desde el inicio congruentes con el corpus que se iba a anotar. Y ese es el trabajo que nos ocupa desde 2015: fue necesario realizar pequeñas adaptaciones en el sistema de etiquetas, como la eliminación de etiquetas para ítems que la transcripción obviaba (símbolos, cifras o fechas), y se requirieron asimismo algunas más en el diccionario (nuevas incorporaciones de formas y nuevas asociaciones de algunas formas preexistentes con etiquetas adicionales), pero donde hubo que llevar a cabo el mayor número de ajustes fue en la aplicación de las etiquetas a las palabras del corpus. De este último aspecto es, precisamente, del que nos ocupamos prioritariamente en este trabajo. En el § 2 se describen algunas facetas relevantes del etiquetador, del etiquetario, del diccionario y del corpus de entrenamiento. En el § 3 se explican algunas particularidades de la etiquetación relacionadas con el carácter oral del corpus ESLORA y en el § 4 se comentan las soluciones adoptadas para resolver algunos problemas de anotación, como los referidos a conectores, fraseologismos, ítems con características peculiares e ítems no reconocidos por el etiquetador.

2. El proceso de anotación

El objetivo final del proyecto ESLORA es la obtención de un corpus oral de español de Galicia que ponga al servicio del usuario consultas que permitan acceder, además de a los datos sobre tipos de interacciones y caracterización sociológica de los interlocutores, a la información de índole lingüística con la que se haya enriquecido el corpus: en este momento, información gramatical sobre clase y subclase de palabras, categorías gramaticales y lema de cada palabra del corpus. Para proveer al corpus de esta información, se ha utilizado un etiquetador que ha requerido la producción de los siguientes recursos de los que se sirve: un etiquetario, incluido el modo en que han de aplicarse sus etiquetas al corpus, un diccionario y un corpus de entrenamiento. A las características de cada uno de estos elementos nos referimos brevemente a continuación[6].

2.1 El etiquetador XIADA

Para añadir información morfosintáctica al corpus ESLORA se ha utilizado el etiquetador XIADA, un etiquetador en principio desarrollado para el gallego y que ha tenido, por tanto, que ser adaptado para el español. XIADA es un

[6] El proceso de anotación descrito en este apartado 2 está reseñado también, aunque con menos detalle, en Barcala *et al*. 2018.

etiquetador fundamentalmente estadístico, aunque permite incluir también reglas lingüísticas que ayudan a mejorar su tasa de acierto (Domínguez 2016). Por ahora el uso de estas reglas no se ha aplicado a la etiquetación de español, aunque está prevista su incorporación en la fase siguiente de mejora de la etiquetación del corpus. Como todos los etiquetadores, XIADA aplica un conjunto de etiquetas, el etiquetario, al corpus objeto de anotación. En tanto que etiquetador estadístico, lo hace fundamentalmente a partir de un diccionario de asociaciones entre formas de palabras y etiquetas, así como de un corpus de entrenamiento que proporciona el modelo de desambiguación. Estos componentes permiten al etiquetador elegir la etiqueta que le corresponde en el contexto en el que se encuentra a cada palabra del corpus que como palabra aislada puede asociarse a más de una etiqueta. Además de por sus satisfactorios resultados, probados para el gallego (Domínguez *et al.* 2009), el etiquetador XIADA fue escogido para la anotación de ESLORA porque, a diferencia de FreeLing, por ejemplo, permite gestionar archivos XML de modo nativo, con lo que podía ser ajustado para manejar conjuntamente la etiquetación morfosintáctica y las marcas de oralidad que se usan en el corpus de forma más eficaz que los otros tenidos en cuenta (Barcala *et al.* 2018). El empleo de XIADA ha permitido que la aplicación de consulta de ESLORA ponga ambos tipos de marcas a disposición del usuario de manera mucho más eficiente y simple.

2.2 El etiquetario[7]

En el momento de la elaboración de este trabajo el etiquetario empleado consta de 453 etiquetas: 198 de pronombres, 136 de determinantes, 78 de verbos, 15 de sustantivos, 13 de adjetivos, 1 de adverbio, 1 de preposición, 1 de conjunción, 1 de interjección, 1 de puntuación y 8 etiquetas para elementos residuales, esto es, símbolos, fechas o cifras, elementos que, en realidad, no aparecen en el corpus, bien por las características del mismo, bien por la transcripción que se ha elegido hacer de tales elementos en los pocos casos en que se utilizan en las entrevistas o conversaciones transcritas.

Desde el punto de vista de su composición o estructura, cada etiqueta está formada por un carácter inicial indicador de categoría seguido donde corresponde de una serie de caracteres indicadores de valores de rasgos gramaticales. En nuestro caso, los valores para la indicación de las subclases de

[7] El conjunto de etiquetas morfosintácticas utilizado en ESLORA no guarda relación alguna con el que XIADA utiliza para el gallego (http://corpus.cirp.gal/xiada/etiquetario/taboa). Es más, difiere de este en el tratamiento de los determinantes y pronombres, entre otros aspectos.

palabras o categorías gramaticales son los siguientes: subclases de sustantivos, determinantes y pronombres, y categorías de género, número, persona, tiempo, modo y caso allí donde son posibles. La disposición, el carácter y la cantidad de información proporcionada en la anotación de ESLORA está, pues, en línea con la tradición de anotación morfosintáctica de corpus en general (representada en primera instancia por el estándar EAGLES) y de corpus del español en particular.

Así, por ejemplo, las etiquetas que en el etiquetario dan cuenta de los determinantes posesivos de primera persona son las siguientes[8]:

DS1AS, Det. Pos. 1.ª-Pers.-Sing. Masc./Neut. Sing., *Es* mío.
DS1EP, Det. Pos. 1.ª-Pers.-Sing. Fem./Masc. Plur., Mis *guardaespaldas me protegerán*.
DS1ES, Det. Pos. 1.ª-Pers.-Sing. Fem./Masc. Sing., Mi *guardaespaldas me protegerá*.
DS1FP, Det. Pos. 1.ª-Pers.-Sing. Fem. Plur., *mías;* Mis *hermanas me ayudarán*.
DS1FS, Det. Pos. 1.ª-Pers.-Sing. Fem. Sing., *mía;* Mi *hermana me ayudará*.
DS1MP, Det. Pos. 1.ª-Pers.-Sing. Masc. Plur., *míos;* Mis *hermanos me ayudarán*.
DS1MS, Det. Pos. 1.ª-Pers.-Sing. Masc. Sing., *Ese es* mío*;* Mi *hermano me ayudará*.
DS1NS, Det. Pos. 1.ª-Pers.-Sing. Neut. Sing., *Eso es* mío.

Como se deduce de la observación de estas etiquetas, el sistema de anotación contempla la asignación contextual de hipervalores de rasgos, es decir, valores que se refieren simultáneamente a más de una especificación para el rasgo en cuestión. Así, los símbolos A o E para el género, que indican respectivamente masculino o neutro y femenino o masculino, son hipervalores de género. Lo peculiar de la presencia de estos hipervalores en nuestro sistema de anotación es que, en los ítems que los admiten, las etiquetas correspondientes coexisten con etiquetas que presentan valores completamente especificados para el rasgo implicado. Además, el sistema prevé que se asignen hipervalores o valores unívocos en razón del contexto, tal como vemos en el ejemplo anterior: *mis* recibe valor de género E (masculino o femenino)

[8] La D es abreviación de Determinante; la S, de la subclase Posesivo; el 1 alude a la persona y al poseedor, primera persona-un solo poseedor; el cuarto carácter indica el género, y el quinto el número.

seguido de *guardaespaldas*, pero valor de género M (masculino) seguido de *hermanos*.

2.3 El diccionario

La tabla 1 muestra las características principales del diccionario utilizado en la etiquetación de ESLORA.

	Asociación forma-etiqueta-lema	Lemas	Formas distintas/formas que entran en alguna ambigüedad
V(erbos)	419 655	5 515	297 627/63 427
A(djetivos)	151 942	22 032	75 223/45 126
N(ombres)	139 030	58 042	125 678/32 285
W(Adverbios)[9]	4 178	4 166	4 178/67
P(ronombres)	833	158	437/419
D(eterminantes)	696	164	435/405
X(Preposiciones)	153	153	153/14
C(onjunciones)	137	137	137/13
I(nterjecciones)	28	28	28/13
Q(Puntuación)	15	15	15/0
Total	716 667	90 410	468 539/106 385

TABLA 1. Atributos del diccionario empleado en la etiquetación de ESLORA

Se trata de un lexicón con un número de lemas amplio (téngase en cuenta que el DLE registra 93 000), de léxico no especializado. La información recogida en la tabla 1 evidencia la importancia de la tarea de la desambiguación, en nuestro caso incrementada por la consideración de hipervalores de rasgos. Vemos que hay 1.5 asociaciones forma-etiqueta-lema por cada forma distinta, y que 1 de cada 4.4 formas es ambigua. La palabra que ofrece mayor ambigüedad, *más*, llega a tener 26 etiquetas posibles.

2.4 El corpus de entrenamiento

El primer paso para poder etiquetar texto en español con XIADA consistió en documentar al menos una vez cada una de las etiquetas presentes en el etiquetario, con el propósito de crear lo que en lingüística de corpus se denomina *Training Corpus Zero* o kérnel, el corpus de entrenamiento cero. Para ello,

[9] El distinto número de formas y lemas de adverbios obedece a los diminutivos, como *carísimamente* o *prontísimo*, recogidos en el lexicón.

se creó un archivo de texto en el que se incluyó, en un formato de forma/etiqueta/lema, separados por tabulaciones, el análisis de 318 enunciados (3102 unidades), separados uno de otro simplemente por una línea en blanco.

Los enunciados, siempre que mostraban el análisis correspondiente a una o varias de las etiquetas cuyo uso se pretendía ilustrar, se extrajeron del propio corpus ESLORA. Cuando esto no era posible por no documentarse en el corpus (por ejemplo, las etiquetas relativas a futuro de subjuntivo), por no ser un uso claro, o por la extensión del fragmento en el que se localizaba la etiqueta, se optó por tomar el enunciado que ejemplifica dicho uso en el propio tagset. Esta tarea se realizó íntegramente a mano, aunque luego ya sí de modo automático se hicieron las comprobaciones pertinentes para cuadrar las etiquetas disponibles en la relación que se deriva del etiquetario y las que se derivan del kérnel.

A continuación se dio inicio al proceso de creación del corpus de entrenamiento propiamente dicho, cuya etiquetación fue ya automática, aunque inicialmente con una tasa de acierto baja por disponer la herramienta únicamente de los datos que proporcionaba el kérnel. Utilizando como corpus de entrenamiento el kérnel, etiquetamos automáticamente una extracción aleatoria de 22 102 palabras de entrevistas de ESLORA y revisamos esa anotación manualmente. Usando después estas 22 102 palabras etiquetadas y revisadas como corpus de entrenamiento etiquetamos de nuevo automáticamente otra extracción aleatoria de fragmentos de ESLORA hasta completar 50 000 palabras, ahora incluyendo conversaciones. Revisamos también este segundo conjunto de palabras y obtuvimos así el corpus de 50 000 palabras que habría de constituir, a partir de entonces, el corpus de entrenamiento.

En cada fase de revisión manual, los revisores trabajaron sobre una salida del etiquetador que conservaba las alternativas de anotación no seleccionadas por la etiquetación automática. Esa salida se integraba en un editor XML[10] personalizado en el que los anotadores podían, de manera simple, aprobar la etiqueta propuesta por la etiquetación automática, seleccionar una de las que habían sido en principio descartadas, o incluso, proponer una que no había sido considerada.

En los siguientes apartados examinaremos las dificultades que tuvieron que afrontar los anotadores en el proceso de revisión y las decisiones que adoptaron. Para facilitar la exposición, hemos agrupado los problemas en dos apartados: a) en el § 3 exponemos las dificultades derivadas del hecho de que el corpus etiquetado presenta marcas específicas de oralidad y está trans-

[10] XMLmind XML Editor, accesible en http://www.xmlmind.com/xmleditor/.

crito siguiendo convenciones específicas; b) en el § 4 revisamos problemas inherentes a la propia etiquetación morfosintáctica: categorización de ítems que, según los casos, son coincidentes con los detectados al etiquetar lengua escrita, parcialmente distintos, o exclusivos de la lengua oral.

3. Etiquetación en relación con la transcripción y las marcas de oralidad presentes en el corpus

3.1 Hipervalores y marcas de oralidad

La presencia de hipervalores para la especificación de los rasgos categoriales exige que se determinen las circunstancias en las que se asignan hipervalores y no valores simples. Las circunstancias en que se asigna, por ejemplo, el valor de género masculino y femenino (E), y no masculino (M) o femenino (F) para el pronombre personal de primera persona en singular *yo*.

En la anotación de ESLORA se asignan hipervalores cuando, en lo que consideramos fragmento o enunciado de referencia, no hay nada que permita determinar un valor simple de entre los posibles para el ítem en cuestión. Más allá de lo que consideramos fragmento, no se tiene en cuenta ninguna información que sea obtenida a partir del discurso anterior o posterior para determinar el valor de una categoría cualquiera. Se le asigna entonces a *yo* el hipervalor de género (masculino o femenino) asociado al carácter E cuando en el fragmento que lo contiene no hay ningún otro elemento que apunte inequívocamente al género masculino (M) o femenino (F) de *yo*.

Ahora bien, para aplicar lo establecido en los párrafos anteriores es preciso determinar qué se entiende por fragmento o enunciado de referencia. En nuestro análisis de ESLORA, el fragmento o enunciado de referencia es exclusivamente la porción de discurso comprendida entre dos pausas largas, dos silencios o una pausa larga y un silencio. Veámoslo con ejemplos[11]: en la figura 1 *yo* recibirá valor M, masculino, porque la presencia de la palabra *pequeño* un poco más adelante permite especificar el género de *yo* como masculino en el fragmento.[12] Sin embargo, en la figura 2 la misma unidad gramatical *yo*, extraída del mismo texto y con el mismo referente, es caracterizada como

[11] Cuando emprendemos la revisión de la etiquetación, visualizamos las marcas de oralidad, al igual que el etiquetador, a través de su representación mediante etiquetas xml en medio de la transcripción, tal como las vemos en los ejemplos a partir de 3. No obstante, en la aplicación de consulta de ESLORA, para solventar el inconveniente de que las etiquetas aparezcan en medio del texto y dificulten su lectura, las palabras afectadas por alguna marca se visualizan sobre fondo amarillo y se activan indicaciones aclaratorias sobre la marca en cuestión al pasar el cursor por alguno de esos elementos.

[12] La existencia de una pausa corta <pausa/> no determina el reconocimiento de un fragmento distinto.

masculino y femenino (E), porque no hay en el fragmento que la contiene nada que permita especificar su género ni como masculino ni como femenino.

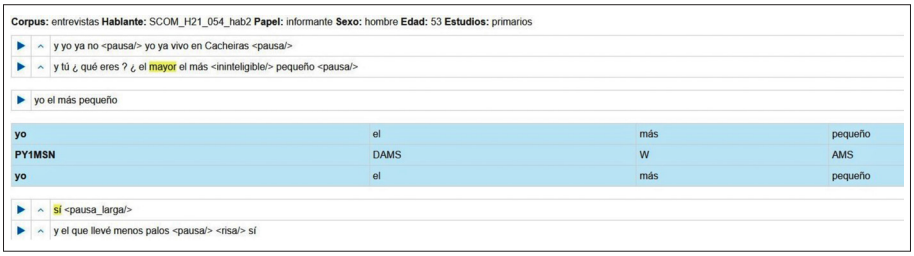


FIGURA 1. Yo en masculino (valor M)

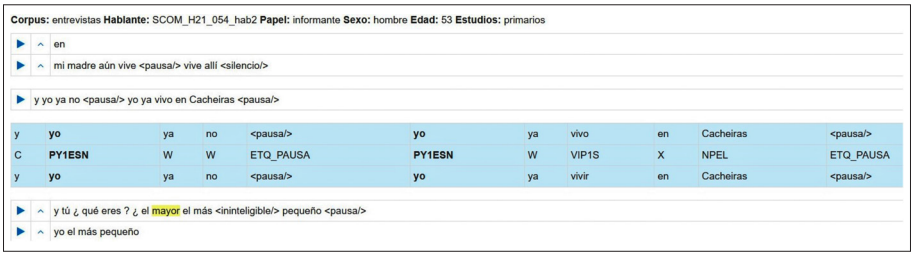


FIGURA 2. Yo en masculino/femenino (hipervalor E)

Junto a los límites dentro de los cuales llevamos a cabo la especificación de valores (el fragmento y no el texto o una unidad gramatical cualquiera), hay que determinar además qué permite en el fragmento establecer la especificación. En nuestro caso, sirve para ello la consonancia de valores reconocida por medio de cualquier tipo de concordancia: a) gramatical con repetición explícita (eco, si se prefiere este término) de valores en las etiquetas de las palabras relacionadas, como en (1); b) gramatical estructural (sin eco de valores), como en (2); c) simple correferencialidad, como en (3), a pesar de que esta última es, casi siempre, contextual y no necesaria.

(1) porque **yo** fui **maestro** <alargamiento>de</alargamiento> de muchas o <alargamiento>sea</alargamiento><pausa_larga/> (SCOM_H31_046)

En (1) existe concordancia gramatical entre el sujeto y el predicativo (atributo) del sujeto, por lo que la etiqueta que corresponde al sustantivo que funciona como predicativo refleja los valores de masculino y singular para el

género y el número y, en razón de ello, el pronombre *yo* los muestra también en su etiqueta.

(2) no tengo **claro qué hacer** <silencio/> (SCOM_H11_051)

En (2) se presenta un caso de concordancia estructural entre el complemento directo (*qué hacer*) y el predicativo del complemento directo (*claro*). El segundo está en neutro singular porque así lo requiere el referente abstracto de su base de predicación: un complemento directo que es una interrogativa indirecta. Sin embargo, esta última no está etiquetada como tal construcción de ninguna manera, por lo que los valores asumidos por el predicativo del complemento directo no pueden estar reproduciendo ningún valor explícito en una etiqueta para esa construcción, no hay aquí posibilidad de eco de valores de una etiqueta a otra.

(3) organizo todo <risa/> <pausa/> yo creo que sí <pausa/> **ellas** no se dan cuenta <pausa/> tendrías que hablar con **ellas** y decír**se**lo no sí que <pausa/> siempre me lo agradecen <pausa_larga/> (SCOM_M13_008)

En (3) ni *ellas* ni *se* concuerdan gramaticalmente, son simplemente correferenciales (aunque no tendrían por qué serlo). En estas circunstancias al pronombre implicado (*se*) siempre se le otorga una etiqueta que lo identifica como corresponde (femenino plural) si el discurso previo avala la correferencia (cuando cabe duda al respecto, se entiende que la avala).

3.2 Etiquetación en relación con las marcas de oralidad

Las etiquetas que se utilizan en ESLORA para dar cuenta de los distintos fenómenos que se producen en el habla son las que se muestran más abajo, repartidas en dos bloques según se trate de etiquetas instantáneas o no, es decir, según presenten una única etiqueta, como las seis primeras, o que engloben texto en su interior, presentando una etiqueta de apertura y otra de cierre, como las siete últimas. A su vez, en los dos bloques debemos distinguir aquellas etiquetas de oralidad que repercuten de algún modo en la etiquetación morfosintáctica, las tres primeras y las cuatro últimas de la lista, agrupadas en (a) y (d), de aquellas otras que no repercuten en la etiquetación morfosintáctica, grupos b) y c).

a) <pausa/>
<pausa_larga/>
<silencio/>

b) <ininteligible/>
<risa/>
<ruido tipo=«xx»/>

c) <sic_inicio/> <sic_fin/>
<lengua_inicio nombre=«xx»/> <lengua_fin/>
<palabra_cortada></palabra_cortada>

d) <alargamiento></alargamiento>
<cita_inicio/><cita_fin/>
<énfasis_inicio/><énfasis_fin/>
<risa_inicio/><risa_fin/>

De las marcas propias de oralidad, establecen pausas en el discurso *pausa*, *pausa_larga* y *silencio*. De ellas, la única que se lematiza y etiqueta es *pausa*. El lema es la marca en sí, y su etiqueta ETQ_PAUSA. Por su parte, *pausa_larga* y *silencio* son la base de la segmentación del turno en fragmentos o enunciados de referencia que, como ya se indicó más arriba, son el ámbito sobre el que operan tanto la etiquetación automática como la manual.

En cuanto a los ítems contenidos en las marcas *palabra_cortada*, *lengua* y *sic*, no se etiquetan ni se lematizan, ni se tiene en cuenta la información que podrían proporcionar para la etiquetación manual de elementos anteriores o posteriores. La razón por la que se procede de esta manera es la siguiente: al tratarse de palabras incompletas, palabras de otras lenguas o errores en la pronunciación, no está al alcance de la etiquetación automática gestionar esa información. Caso distinto es el de los ítems comprendidos entre las marcas *cita*, *alargamiento* o *énfasis*. Las palabras por ellas abarcadas se analizan igual que las que no están afectadas por ninguna marca, proporcionándoles el lema y la etiqueta que les correspondan conforme al contexto.

En (4) y (5) podemos ver algunos ejemplos de estas marcas y de las anteriormente mencionadas. Los colores reflejan las siguientes convenciones: en rojo aparecen las marcas que engloban texto que no se etiqueta ni se tiene en cuenta en la etiquetación, por ejemplo, *palabra_cortada* (*si* en el primer segmento) o *lengua*, en este caso gallego (el fragmento *mira*, *aquí vai o o colexio e aquí vai o centro sociocultural e aquí van as piscinas*). En verde se presentan aquellas otras marcas que comprenden texto analizable al que se le añade como información adicional la que proporciona la marca de oral en cuestión; por ejemplo, para *no* en (4) y para *o* y *sociocultural* en (5), se añade la información de que son formas que presentan un alargamiento en su realización, o para *bueno mira aquí voy a* la de que reproduce una cita. Por

último, la única marca a la que se le asigna una etiqueta y un lema aparece en los ejemplos en color azul.

(4) no <palabra_cortada>si</palabra_cortada> no sirve de nada <alargamiento>no</alargamiento><pausa/> (SCOM_H11_051)

(5) y entonces ¿qué ocurría? pues ocurrió lo siguiente que llegábamos nosotros y decías <cita_inicio/>bueno mira <pausa/> aquí voy a<cita_fin/> llegaba al ayuntamiento ¿no? el equipo de gobierno que era socialista <cita_inicio/><lengua_inicio nombre=«gl»/>mira aquí vai <pausa/><alargamiento>o</alargamiento><pausa/> o colexio <pausa/> e aquí vai o centro <alargamiento>sociocultural</alargamiento> e aquí van as piscinas<lengua_fin/><cita_fin/><pausa/> pues llegábamos nosotros y decía <cita_inicio/>no mira<cita_fin/> (SCOM_H33_002)

Un problema adicional en relación con las marcas de oralidad viene determinado por la herramienta de etiquetación automática. Esta exige que las unidades a las que se les va a asignar una etiqueta, sean estas simples o multipalabra, no puedan estar interrumpidas por las marcas de oralidad en el texto. Por esa razón, por ejemplo, se decidió en la transcripción que el alargamiento o el énfasis afecten a palabras completas y no solo a alguna de sus sílabas, lo cual habría podido constituir un reflejo más fiel de lo que en realidad ocurre. Por otro lado, la transcripción no tiene a su alcance, ni se ocupa de ese reconocimiento, la información que identifica a series de palabras como unidades multipalabra, por lo que marcas como las referidas, si afectan a una palabra integrante de una unidad multipalabra, aparecen en el corpus abarcando solo a esa palabra una vez que el texto se somete al proceso de etiquetación. La unidad multipalabra se encuentra, así, interrumpida en el texto sujeto a etiquetación y el etiquetador automático no será capaz de identificarla. En la revisión manual, no obstante, ante una unidad multipalabra que ofrezca estas características en relación con determinadas marcas (por ejemplo, con alargamiento en alguna de las palabras que constituyen la unidad multipalabra, *de* en *dentro <alargamiento>de</alargamiento> tres o cuatro años*), se modifica la transcripción para que la información aportada por la marca de oralidad afecte a la unidad completa, por lo que el ejemplo anterior se convierte en *<alargamiento>dentro de</alargamiento> tres o cuatro años*. Pero está claro que estos cambios solo serán posibles en el fragmento del corpus manualmente revisado.

No se procede así, de todos modos, con todas las marcas de oralidad que pueden romper las unidades multipalabra. Así, la *pausa* o la *pausa_larga*,

que pueden llegar a fragmentar la unidad multipalabra en dos enunciados de referencia sucesivos, no son modificadas para reagrupar la unidad multipalabra, porque se considera que ello alteraría más de lo razonable la transcripción. En estos casos, lo que se hace es asignarles a ambas partes de la unidad multipalabra fragmentada o discontinua la misma etiqueta: así, en el ejemplo (6), PCMS (pronombre numeral cardinal masculino singular) es asignada a *mil*, primero, y a *novecientos cuarenta*, a continuación.

(6) es decir de estar <pausa/> un año con mis tíos <pausa/> fue cuando yo hice la primera comunión <pausa/> el día <pausa/> siete el <palabra_cortada>dí</palabra_cortada> el día dieciséis de junio de **mil** <pausa/> **novecientos cuarenta** <pausa/> [...] (SCOM_H33_015)

Por otra parte, en el registro escrito la concurrencia sucesiva de una misma forma ortográfica suele corresponder a unidades sintácticas diferentes (*Pónse*la la *noche antes de iros*), salvo en caso de erratas; en cambio, la repetición consecutiva de formas es uno de los fenómenos característicos de la oralidad. Frente a la solución adoptada en otros corpus, como el COSER, esta peculiaridad no se codifica en ESLORA de ningún modo específico, ni en el nivel de transcripción ni en el de anotación morfosintáctica. Simplemente se transcriben y se etiquetan tantas formas idénticas como emita el hablante. La figura 3 lo ilustra:

FIGURA 3. Muestra del análisis de dos formas repetidas del artículo femenino

Respecto a las no finalizaciones, hemos visto más arriba que en ESLORA se creó una etiqueta específica para marcar las palabras cortadas, que el etiquetador ignora; sin embargo, en el nivel superior a la palabra, no se codifican las frases o secuencias inacabadas. Para ejemplificar ambos casos reproducimos en la figura 4 el análisis que proporciona la aplicación de consulta. Repárese en que *siem*, para la que no se proporciona ni etiqueta ni lema en el análisis morfológico, se visualiza en amarillo en la secuencia y al

situar encima el ratón emerge el texto señalando que estamos ante una palabra cortada, codificación que procede de la transcripción. Mientras que, por otra parte, la marca amarilla de la preposición final corresponde a un alargamiento y no, como podría interpretarse, a una secuencia inacabada:

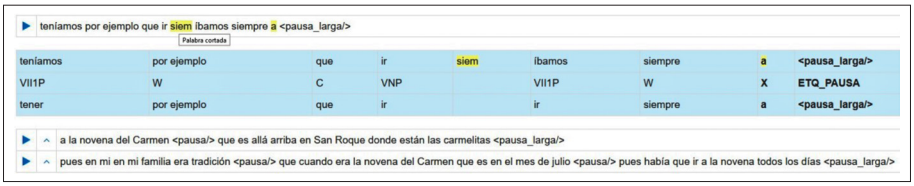


Figura 4. Muestra del análisis de una secuencia inacabada

Por último, aunque sí se codifican los solapamientos en el nivel de transcripción, en el proceso de construcción del corpus de entrenamiento se prescindió de esa información. Así, los segmentos extraídos aleatoriamente para desambiguar no incluían ya información sobre superposiciones, pues si bien la mera existencia de texto solapado implica la existencia de otro hablante, lo cual resulta fácilmente representable e interpretable en ELAN, estamos todavía trabajando en cómo manejar y trasladar esa información a la aplicación de consulta.

3.3 Etiquetación en relación con la transcripción

No podemos entrar aquí en los detalles de la transcripción ortográfica de ESLORA, pero algunas de las decisiones tomadas para llevarla a cabo, como es esperable, tienen una repercusión importante en el proceso de anotación. Mencionamos a continuación la respuesta que hemos dado en la anotación a algunas de las decisiones de la transcripción que podían condicionarla.

En primer lugar, nos referiremos a la etiquetación de los numerales. En ESLORA, tal y como recomienda EAGLES, los numerales se transliteran siempre como palabras, esto es, con caracteres alfabéticos, nunca numéricos. Respecto a su etiquetación, a semejanza de lo decidido en CORPES XXI y frente a otros corpus escritos que clasifican las expresiones numéricas bajo las etiquetas genéricas de Cifra, Fecha y Hora (CAES y AnCora), en ESLORA etiquetamos los numerales como determinantes o pronombres numerales, simples o multipalabra, en función de cada caso concreto.

Por el momento no está implementado el módulo de reconocimiento de numerales multipalabra, por lo que la salida del etiquetador no identifica los

numerales constituidos por más de una unidad ortográfica, de modo que el revisor debe reconstruir la unidad multipalabra correspondiente a los numerales donde la identifique y asignar la etiqueta correspondiente. En general, la anotación morfosintáctica del numeral se lleva a cabo de acuerdo con las pautas indicadas en los siguientes apartados:

a) Cantidades/cifras

Se caracterizan como determinantes o pronombres cardinales, según precedan a sustantivo o no, DC/PC, singular para el *token uno/una*, plural para los demás. Su género concuerda con el del sustantivo al que se refieren. Así determinante, DCMP, para *cincuenta y dos* en (7), frente a pronombre —PCMP— en (8) y (9), donde además se ha decidido que *y tantos*, *y pico* formen parte del numeral multipalabra, analizando en consecuencia.

(7) hace **cincuenta y dos** años <silencio/> (SCOM_H21_054)

(8) pues tendría no sé **cuarenta y tantos** casi **cincuenta** <pausa_larga/> (SCOM_H32_032)

(9) sí sí <pausa/> anda sobre **ciento y pico** de euros <pausa_larga/> (SCOM_H31_043)

b) Años y días del mes

Los años y días se caracterizan como pronombre cardinal masculino singular: PCMS. Las fechas se descomponen en sus elementos constituyentes, como se puede ver en el análisis reproducido en (10) para *dieciocho de enero de dos mil seis*.

(10) fue <pausa/> el **dieciocho de enero de dos mil seis** <pausa_larga/> (SCOM_H33_007)
— dieciocho: PCMS/dieciocho
— de: X/de
— enero: NCMS/enero
— de: X/de
— dos mil seis: PCMS/ dos mil seis

c) Horas

No contemplamos una etiqueta *hora*, como ya se indicó, por lo que la indicación horaria se segmenta en sus partes constitutivas. El numeral se etiqueta como pronombre cardinal femenino plural, salvo *una*, que es singular. Si no se indica hora justa, sino con *y cuarto* o *y media*, analizamos por sepa-

rado cada elemento, como en el ejemplo de (11): en primer lugar el numeral como pronombre cardinal femenino plural y a continuación la conjunción *y* seguida del pronombre partitivo *cuarto* o *media*:

(11) mm estaría ya bien pero yo a las **ocho ocho y cuarto** como <énfasis_inicio/> muy <énfasis_fin/> tarde ya estoy en el colegio <pausa_larga/> (SCOM_M23_018)

d) Porcentajes

Los porcentajes constan del número, como *ocho*, etiquetado como pronombre cardinal masculino singular, y el adverbio multipalabra *por ciento*. No es infrecuente aquí, además, que el numeral no sea entero. En casos como estos (12), donde el porcentaje 8,5 se puede expresar como *ocho con cinco / ocho y medio / ocho cincuenta*, decidimos separar cada uno de los componentes.

(12) y es ese **ocho y medio por ciento** vendas una cajetilla vendas diez millones <pausa/> (SCOM_H23_003)
— ocho: PCMS/ocho
— y: C/y
— medio: PPMS/medio
— por ciento: W/por ciento

Por último, otro de los aspectos que ha suscitado dudas es el de los múltiples ruidos comunicativos, onomatopeyas e interjecciones no lingüísticas que, a no ser que se trate de los ítems más estandarizados, se presentan de un modo heterogéneo en la transcripción, la cual ofrece varias formas de transliteración, con más o menos consonantes/vocales para representar ese sonido no lingüístico (por ejemplo *mmm*, *mm*, *mmmm*, *mmn*, *hmmm*, *mhum*, *mhmmm*, *mh* o también *aaah*, *aahh*, *aaaah*, *aamm*, *ahm*, *ah*, *ahh*, *ahhhh*).

Como se ha señalado, el texto llega a los analistas ya transcrito y, si bien convenimos en etiquetar todos estos ítems como interjecciones, reconocemos también la necesidad de una mayor homogeneización en algunos de ellos, mejora que actualmente los responsables de la transcripción ya están introduciendo, por lo que en ESLORA este será un problema que dejará de plantearse ya a partir de la versión 1.2.2.

Muchos de los ruidos comunicativos u onomatopeyas anteriores son formas desconocidas para el etiquetador, puesto que no aparecen en el lexicón o diccionario que utiliza. El etiquetador debe recurrir, por tanto, al módulo de adivinación y, en función de la terminación de la palabra desconocida y de la

información sobre los elementos anteriores y posteriores, debe aventurar un análisis, que no siempre es acertado. A modo de muestra, hemos contrastado la etiquetación de *bo* y *boh*, por una parte, y de *fff*, *ff*, y *eehmm*, por otra, en la versión 1.2.2 de ESLORA (noviembre 2018). Los resultados difieren: XIADA nunca etiqueta correctamente como interjección *bo y boh*, pero acierta plenamente con los 202 casos de *fff*, los 138 de *ff*, o los 73 de *eehmm*.

4. Problemas de etiquetación

4.1 Cuestiones básicas

Algunos aspectos relacionados con la anotación de determinados elementos problemáticos ya eran previsibles antes de dar comienzo al trabajo con ESLORA, bien por nuestra experiencia previa en etiquetación de texto escrito, bien por nuestro conocimiento de usos propios del español en general o de la variedad concreta que analizamos. Así, por ejemplo, antes de empezar a etiquetar, ya sabíamos que tendríamos que tomar una decisión sobre cómo habrían de ser tratados los tiempos compuestos y las perífrasis verbales. Para estas unidades, en ESLORA no consideramos conveniente ni introducir en el lexicón todas las formas de tiempos compuestos para cada uno de los verbos ni diseñar una estrategia de posprocesado para identificarlos, de modo que los tiempos compuestos y las perífrasis son etiquetados separando auxiliar de auxiliado: *he cantado* se etiqueta como VIP1S VPMS y *vamos a ir*, como VIP1P X VNP.

Otro caso previsible era la aparición del pronombre *le* con referente plural (13). De nuevo, se optó por la asignación del valor en razón únicamente de la forma.

(13) a nosotros rama **le** llamamos a **a los ramitos** que tenía la la que sale del del del tronco grande ¿ no sabes ? aquellas pequeñitas y toda la cosa <pausa/> [...] (SCOM_M33_009)

Por otra parte, también preveíamos que aparecerían en ESLORA, al igual que en español general, las formas en *-ar/-er/-ir* para expresar la modalidad exhortativa, como en (14). De nuevo, se decidió etiquetar estas formas conforme a su valor canónico, es decir, como infinitivos.

(14) [...] él nos dijo <cita_inicio/>¿ vais a Estambul ? <pausa/> pues en Estambul hay un tío en el eh <pausa/> el director del Cervantes de Estambul <pausa/> es un un un rapaz también del Pais Vasco <pausa/> que yo lo conozco <pausa/> **poneros** en contacto con él a través de mí o sea <pausa/> **decirle**

que que tal <cita_fin/><pausa/> y entonces <pausa/> a través de Marquiegui <silencio/> (SCOM_M32_025)

En la misma línea, dado que ESLORA es un corpus que registra los usos orales del español de Galicia, sabíamos también desde el comienzo que era necesario contar con que las formas verbales *-ra* son usadas con valor de antepretérito (15), valor codificado en la norma del español europeo por medio de la combinación *había+participio*. En la etiquetación del corpus se decidió que no se reflejaría el valor en cuestión, por lo que se optó por codificar como VSI* (pretérito de subjuntivo) todas las formas en *-ra*. Habrá de ser, pues, la desambiguación manual, una vez realizada la búsqueda en el corpus, la que permitirá a los estudiosos de las formas verbales del español de Galicia identificar el valor modotemporal específico.

(15) [...] claro <pausa/> yo empecé a trabajar <pausa/> en agosto el año pasado <pausa/> y fue un mes <pausa/> mi marido **cogiera** vacaciones quince días <pausa_larga/> (SCOM_M11_040)

4.2 Los conectores

Los conectores discursivos[13] son también unidades lingüísticas especialmente problemáticas para la categorización. No queriendo asignarles una etiqueta específica de conectores, determinar para ellos una etiqueta gramatical no es fácil: los diccionarios de uso general o no la proporcionan o son inconsistentes al respecto y los estudios que se ocupan de ellos los estudian en su función discursiva más que desde el punto de vista de su caracterización gramatical. Martín Zorraquino & Portolés Lázaro (1999: 4016) dicen que «se ajustan en general, a las categorías tradicionales de los adverbios, de las locuciones adverbiales y de ciertas interjecciones», pero, al margen de ello, solo en muy contados casos se refieren a las unidades que revisan por medio de caracterizaciones gramaticales, utilizando más bien denominaciones generales como *unidad* o *marcador*. Tanto Portolés (2001) como Martín Zorraquino (2010) son más explícitos en este sentido, e incluso revisan caracterizaciones previamente atribuidas en Martín Zorraquino & Portolés

[13] No podemos ofrecer aquí una discusión sobre el concepto de marcador/conector del discurso, ni sobre la idoneidad de su aplicación a unidades cuya función es guiar las inferencias que se realizan en la comunicación, en consonancia con sus propiedades específicas, morfosintácticas, semánticas y pragmáticas (Martín Zorraquino & Portolés Lázaro 1999). En el elenco de unidades así catalogadas pueden entrar tanto los elementos mencionados en este apartado como algunos del § 4.3. Su inclusión en epígrafes distintos obedece a que los primeros tienen una función más propiamente conectiva y los segundos una función más evaluativa o conversacional.

Lázaro (1999) (así sucede con respecto a *pues*, reconocido en 1999 como adverbio en su uso discursivo y en las obras posteriores como posible conjunción, en el caso de Portolés, y como conjunción en el de Zorraquino), pero aun así, sus propuestas no son siempre fáciles de aplicar en casos concretos y con frecuencia permanecen las dudas especialmente con respecto a lo que pueda llegar a considerarse interjección.

Por supuesto, una parte importante del problema tiene que ver con el hecho de que algunos conectores están inmersos en un proceso de gramaticalización (*en plan*, *es que*) y/o predominan en la oralidad (*y eso que*, *total*, *nada*) (ambas dificultades son precisamente señaladas por Pons (2000) al adoptar una perspectiva no discreta para el análisis de los marcadores del discurso y al referirse a la falta de un marco teórico en la gramática tradicional con respecto a los fenómenos de la conversación). Así, intentando tomar como referencia el diccionario de la RAE nos encontramos que, aunque para algunos marcadores de esta clase sí se nos ofrecen soluciones que hemos considerado aceptables (conjunciones, *y eso que*, o adverbios, *sobre todo*, *de hecho*, *así y todo*, *al fin y al cabo*), para muchos otros no: unos, como *si cuadra*, *por otro lado*, *a su vez*, no están recogidos en el diccionario, y otros, como *a ver*, *o sea*, reciben una caracterización demasiado amplia y ambigua (*expresión*), no válida para la etiquetación morfosintáctica.

Finalmente, a la difícil categorización de estos elementos se suma en ocasiones el problema de la determinación de su extensión. En general, en el caso más común, si al eliminar una preposición o una conjunción final candidata a integrar el conector, la unidad resultante constituye un enunciado viable, se interpretará como una unidad multipalabra incrementada con alguna clase de modificación y, por tanto, otorgaremos dos etiquetas (*al lado de*: W X, *al margen de*: W X). En caso contrario, la etiquetación será conjunta (*en caso de*: X).

4.3 Frases hechas, fórmulas estereotipadas, frases proverbiales, refranes, etc.

Junto a las anteriores, son también unidades problemáticas otras que, según los autores, a veces se recogen entre los marcadores discursivos y a veces no (Martín Zorraquino 2010: 102-104 aborda esta cuestión). Son unidades de distintos tipos constitutivamente hablando (palabras simples, frases hechas, fórmulas estereotipadas o rutinarias de diversa índole) y sirven para la expresión, entre otros valores, de confirmación (*claro*, *vale*, *de acuerdo*), contraposición (*no sé*, *para nada*), sorpresa (*hala*, *ostra*, *qué va*, *madre mía*, *Jesús*, *Dios*, *Dios mío*), saludo (*hola*, *buenas tardes*), reformulación (*digamos*) o apelación (*venga*, *vamos*, *hombre*, *mujer*, *tío*, *hijo mío*).

Elementos de esta clase como *buenos días*, *hola*, *hala* o *buenas tardes*, por su uso rutinario, plantean, en principio, menos dudas y se ha decidido etiquetarlos como interjecciones, es decir, como I[14]. Lo mismo sucede con *madre mía*, que aparece en ESLORA como lema y una expresión de la que se obtienen 25 casos como interjección, o *Dios mío*, expresión de la que existen en ESLORA 65 casos para etiquetar del mismo modo. Sin embargo, hemos comprobado que muchos de estos ejemplos (*buenos días* o *madre mía*) se etiquetan siempre desagregadamente en CORPES XXI, como combinaciones regulares de sustantivos y adjetivos, probablemente porque, como ya indicamos más arriba, no se ha diferenciado netamente entre la etiquetación de los textos escritos, cuyo modelo ha predominado, y los orales.

Al lado de estos elementos más claros aparecen, no obstante, otros que no lo son tanto. Estos casos, bien porque no son frecuentes en el registro escrito o bien porque no lo son al menos en el uso que tienen en el oral, no han sido homogéneamente descritos desde el punto de vista gramatical al menos en las obras de más larga tradición. En otras más recientes, como la NGLE (2009: 627), se menciona precisamente el desacuerdo entre gramáticos al considerar alguna de estas unidades como interjecciones o no, de modo que en estos casos se torna, en efecto, necesaria una toma de decisiones que resulta más controvertida. Es lo que sucede con *hombre* que, frente a su uso predominante como sustantivo en textos escritos (el único que registra el CORPES), funciona como marcador discursivo[15] en la oralidad, y en estos casos es etiquetado como interjección en ESLORA (coincidiendo con la NGLE (2009) y Briz (2012). Así, de 575 casos de *hombre* en nuestro corpus, solo 32 reciben la etiqueta de N (sustantivo), frente a los otros 543, el 94,43 % de las apariciones, en los que es etiquetado como I. El empleo de *hombre* como interjección aparece recogido en el DLE en la acepción 8, aunque solo se señalan algunos de los valores posibles que encontramos en el corpus. Un ejemplo de uso lo vemos en (16).

(16) <ruido tipo=«chasquido boca»/> **hombre** a ver es ff <palabra_cortada>jod</palabra_cortada> mmm los niños con doce años <pausa_larga/> (SCOM_H11_052)

[14] En la versión 1.2.2 del corpus no están recogidas estas consideraciones en todos los casos.

[15] Destacamos aquí el uso de este término (MD) como etiqueta morfosintáctica en C-ORAL-ROM español, que no es compartido por el grupo que desarrolla el corpus equivalente para el francés (Guirao y Moreno 2006). Entendemos en ESLORA que marcador discursivo es una función pragmática y debe ser diferenciada de la clase de palabras que la sustenta. Por otra parte, nos cuesta entender cómo se distingue entre las etiquetas de marcador discursivo e interjección, tal y como se aplican en C-ORAL-ROM.

Otro caso llamativo es el de *digamos*. En el DLE aparece con la etiqueta de *expresión coloquial*, en un aparte dentro de la entrada del verbo *decir*. En ESLORA se registra 169 veces, en la mayor parte de las cuales está empleado como reformulador o aproximador, usos en los que por ahora ha sido etiquetado como VMP1P, es decir, como forma verbal conjugada. Tomamos esta decisión (provisional y conservadora) en virtud de la dificultad para determinar el grado de gramaticalización del elemento en cuestión, pues no vemos tan claramente que *digamos* haya perdido su significado como forma verbal (ejemplo 17). Lo mismo sucede con *no sé* o con *qué va*, por ahora no lematizados como tales, pero cuyo uso como marcadores resulta asimismo frecuente en el corpus. También en razón de su gramaticalización considera la NGLE (2009: 627) que estas formas deben ser consideradas interjecciones o no.

(17) de mi clase social porque aunque yo no **digamos** no he tenido un <pausa/> trabajo <pausa/> nunca en plan <pausa/> <alargamiento>de</alargamiento> <pausa_larga/> (SCOM_H11_052)

De forma general, pues, diremos que las unidades multipalabra recurrentes de esta clase tienden a etiquetarse en ESLORA como interjecciones, aunque no siempre (*no sé*). Si son unidades simples, se etiquetan conforme a su valor como no marcadores, siempre que su uso como tales no contradiga su significado (así entendemos que sucede con *digamos*). Se etiquetan como interjecciones cuando el significado que les correspondería es incompatible con su valor en el corpus, lo cual sucede con *hombre*, que es utilizado como marcador también cuando la interlocutora es una mujer —lo mismo propone Briz (2012) para *hombre*—.

Por otra parte, además de lo difícil que es determinar la etiqueta o si la etiquetación debe ser conjunta o no, a veces se presenta también el problema de la delimitación del marcador multipalabra de manera similar a como se presenta para los conectores (*de verdad/de verdad que*, *menos mal/menos mal que*), por lo que la decisión tomada para este tipo de combinaciones coincide con la adoptada para las expresiones multipalabra, tanto en el caso ya comentado de los conectores como en otros: locuciones prepositivas, conjuntivas y adverbiales de función oracional.

4.4 Algunos ítems en concreto

Mención especial merecen algunos ítems específicos especialmente dificultosos, como *que* o *tal*. Por lo general, las categorías de *que* (relativo o conjunción) son fácilmente delimitables en la lengua escrita, pero no así en la oral;

con frecuencia aquí *que* introduce apartes que sirven para contrarrestar la información previa y no se manifiesta claramente ni en relación con un posible antecedente ni al margen de él. Se ha decidido etiquetar *que* como relativo (PL) cuando su antecedente esté dentro del fragmento y como conjunción (C) cuando aquel figure fuera del mismo, cuando sea dudoso o cuando no lo haya. Así, como en el ejemplo de (18) existe la duda de que la forma *que* remita a *la gaviota*, se ha considerado conjunción.

(18) de una vez se rompió una rueda y de la otra vez se <alargamiento>rompió</alargamiento> un poco más todo <pausa/><risa/> la parte de adelante y así <pausa/> y el parabrisas <pausa/> porque la gaviota <alargamiento>estaba</alargamiento><pausa/> fuerte eh <pausa/><risa/> y me dio en el parabrisas y rompió el parabrisas eh <pausa/> **que** <alargamiento>estaba</alargamiento> potente <pausa/> esa comía bien ¿ves? <pausa_larga/> (SCOM_H11_047)

En cuanto a la forma *tal*, el DLE la describe como un adjetivo demostrativo o indefinido, o como un pronombre demostrativo. Para dar cuenta de esos usos de *tal*, son adecuadas las etiquetas de DD?? y PD?? de nuestro sistema de anotación, especificadas con los diferentes caracteres para dar cuenta de género y número en la tercera y cuarta posiciones. Pero si la clasificación del DLE es idónea para los usos propios de la lengua escrita, en la lengua oral son particularmente frecuentes usos como los de (19) y (20), que podríamos calificar de discurso de relleno:

(19) y tercero sí <pausa/> en una productora en tercero <alargamiento>y</alargamiento> entonces tuve la suerte de ff de ponerme las pilas rápido <pausa/> empecé a hacer cortos <pausa/> trabajaba en la productora en la productora aprendía veía cómo hacían las cosas y luego por nuestra cuenta íbamos haciendo cortos y **tal** <pausa/> <ruido tipo=«inspiración»/><pausa_larga/> (SCOM_H13_014)

(20) a veces tiene los dos efectos <pausa/> una cara de dos monedas <pausa/> **tal** de <pausa/> desinhibidor social sí pero de factor depresivo <pausa/> es un factor depresivo severo no sé quién decía <pausa/> no sé si era <pausa_larga/> (SCOM_H13_012)

De hecho, si se lleva a cabo una búsqueda rápida de *tal* en ESLORA, se observa que la primera página de resultados no proporciona ni uno solo de los usos previstos en el DLE, sino los ilustrados en (19) y (20). La etiqueta finalmente adoptada para *tal* en estos usos orales, no previstos por la gramática ni el diccionario, ha sido PDIS (que correspondería a pronombre, demos-

trativo, masculino, femenino o neutro, singular), etiqueta que difícilmente se asignará a *tal* en ninguna otra circunstancia.

4.5 Ítems no reconocidos

Además de las onomatopeyas y múltiples ruidos comunicativos, ya mencionados en el apartado 3.3, otros elementos no reconocidos para el etiquetador son formas acortadas, galleguismos, extranjerismos e ítems con sufijos apreciativos y superlativos. Una característica común a todos estos ítems es que no presentan lema, con las dificultades que ello entraña para la recuperación de información del corpus, por lo que actualmente estamos trabajando en el desarrollo de estrategias que permitan lematizarlos, si no en su totalidad, al menos en parte.

En cuanto a los acortamientos o formas a las que se les suprime la parte final, como *profe*, *quimio* o *peli*, decidimos remitirlos al lema con forma plena: *profesor*, *quimioterapia* o *película*, respectivamente. El alto índice de error en la etiquetación de estos elementos, como los 14 casos de *peli/pelis* o los 9 de *quimio*, es decir, todas las ocurrencias de estos elementos, se soluciona incluyéndolos en el lexicón.

De otra parte, están todas aquellas formas verbales que escapan de la variante considerada estándar, como por ejemplo los pretéritos de segunda singular en –s (*dijistes*), o los pretéritos regularizados tipo *reducieron*, o las formas de presente de subjuntivo influenciadas de un modo evidente por el gallego como *estea*. Todos ellos reciben la etiqueta que corresponde: pretérito segunda singular en el primer caso, pretérito tercera plural en el segundo caso y presente de subjuntivo de *estar* en el tercero. De los 28 casos de pretérito en –*stes* que se registran en la versión 1.2.2 de ESLORA, ninguno es etiquetado correctamente, aunque los 5 casos que existen de *vistes* sí se identifican como formas verbales, pero no de *ver*, sino de *vestir*, de ahí que estemos considerando la posibilidad de incorporar en el sistema verbal algunas de estas formas.

Respecto a los galleguismos y extranjerismos que la transcripción no identifica con la marca de *lengua*, los únicos de los que la etiquetación tiene que dar cuenta, no hay un patrón. La tasa de acierto está en función de la terminación y el contexto de cada caso concreto, y así, frente al acierto de los 13 casos de *estornela* como NCFS, está el error para las dos ocurrencias de *smartphone*, que se identifican como forma verbal porque la terminación -*one* aparece con frecuencia en el corpus en otras 123 ocasiones en formas verbales, indicativas, subjuntivas o imperativas, formas como *gestione*, *impone*, *perdone*, *solucione*, *pone*, *supone*, etc.

5. Conclusiones

En esta contribución hemos querido mostrar las razones por las cuales el corpus ESLORA se ha anotado a partir de recursos adaptados específicamente para él, derivadas del hecho de que se trata de un corpus oral del español de Galicia. A partir de la experiencia previa, tanto propia como ajena, habíamos constatado que, para la etiquetación estadística, era necesario contar con un corpus de entrenamiento que fuera también de lengua oral. La producción manual de este corpus de entrenamiento para la etiquetación de ESLORA ha permitido identificar numerosos problemas de anotación morfosintáctica, que tienen su origen en el tipo de variedad que se analiza: la lengua oral.

Para su presentación en este trabajo, hemos agrupado estos problemas en dos apartados. En el primero (§ 3) se han reunido los derivados de la transcripción y de la codificación de los rasgos de oralidad, para los que la etiquetación debe proporcionar soluciones específicas. En el segundo (§ 4) se han recopilado algunos problemas inherentes a la propia etiquetación morfosintáctica, además de los que afectan a la delimitación y categorización de formas y expresiones propias o exclusivas de la oralidad. Con respecto a estos últimos, es preciso destacar que, durante el proceso de desambiguación manual de un corpus, sea escrito u oral, es siempre necesario tomar decisiones para resolver problemas de distinta naturaleza, muchos de los cuales no han sido extensiva u homogéneamente contemplados y analizados en estudios previos. El proceso de toma de decisiones es más complejo cuando se trabaja con corpus orales, por las dificultades inherentes a este tipo de variedad y por la presencia de elementos que no pueden ser catalogados conforme a los estándares previstos para la lengua escrita. Ello implica que los anotadores se tienen que enfrentar a nuevas dificultades para cuya etiquetación es solo parcialmente reutilizable su experiencia previa en este terreno. Esto sucede especialmente cuando hay que anotar elementos que funcionan como marcadores discursivos, o como conectores, o con ciertos ítems (*que*, *tal*) cuyo comportamiento en la lengua oral es muy distinto del que ofrecen en la lengua escrita.

Los ejemplos con los que se han ilustrado en este trabajo los problemas de anotación morfosintáctica también ponen de relieve las ventajas que proporciona dicha anotación con vistas al estudio de numerosos fenómenos lingüísticos a partir de datos de corpus. Con objeto de rentabilizar la etiquetación, consideramos necesario poner a disposición de los usuarios, en las propias páginas de acceso a los datos, la documentación que la explica: etiquetarios y manuales de aplicación de las etiquetas, cuanto más completos, mejor.

Recursos electrónicos

AnCora: http://clic.ub.edu/corpus/es/ancora [consultado: 16/1/2019].

CAES: Instituto Cervantes: *Corpus de aprendices de español como lengua extranjera* (CAES). http://galvan.usc.es/caes/ [consultado: 16/01/2019].

C-ORAL-ROM: http://www.lllf.uam.es/ESP/Coralrom.html [consultado: 5/4/2019].

C-ORAL-BRASIL: http://www.c-oral-brasil.org/ [consultado: 5/4/2019].

CORPES XXI: Real Academia Española: Banco de datos (CORPES XXI) [en línea]. *Corpus del Español del Siglo XXI (CORPES)*. http://www.rae.es [consultado: 16/01/2019]

COSER: http://www.corpusrural.es/ *Corpus oral y sonoro del español rural (COSER)* [consultado: 5/4/2019].

EAGLES (1996): *Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Document EAG–TCWG–MAC/R. http://www.ilc.cnr.it/EAGLES96/browse.html [consultado: 20/6/2018].

EAGLES (1996): *Preliminary recommendations on Spoken Texts*. EAG–TCWG–SPT/P. Version of May, 1996. http://www.ilc.cnr.it/EAGLES96/spokentx/spokentx.html [consultado: 16/01/2019].

ESLORA: http://eslora.usc.es/ *Corpus para el estudio del español oral*, versión 1.2.2 de noviembre de 2018, ISSN: 2444-1430 [consultado: 8/3/2019].

FreeLing: http://nlp.lsi.upc.edu/freeling/index.php [Consultado: 5/4/2016].

Real Academia Española: *Diccionario de la lengua española*, 23.ª ed., [versión 23.2 en línea]. https://dle.rae.es [consultado: 20/6/2019].

Val.Es.CO: Cabedo, Adrián & Salvador Pons (eds.): *Corpus Val.Es.Co 2.0*. http://www.valesco.es [consultado: 5/4/2016].

XIADA: Centro Ramón Piñeiro para a investigación en humanidades http://corpus.cirp.gal/xiada

Referencias bibliográficas

Barcala Rodríguez, Fco. Mario, Eva M.ª Domínguez Noya, Alba Fernández Rodríguez, Raquel Rivas Cabanelas, M.ª Paula Santalla del Río, Victoria Vázquez Rozas & Rebeca Villapol Baltar (2018): «El corpus ESLORA de español oral: diseño, desarrollo y explotación», *CHIMERA. Romance Corpora and Linguistic Studies* 5/2, pp. 217-237. https://doi.org/10.15366/chimera2018.5.2.003

Bick, Eckhard, Heliana Mello, Alessandro Panunzi & Tommaso Raso (2012): «The annotation of the C-ORAL-BRASIL spoken corpus using an adaptation of the Palavras Parser», in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC2012)*. Paris: ELRA, pp. 3382-3386.

Briz Gómez, Antonio (2012): «La definición de las partículas discursivas *hombre* y *mujer*», *Anuario de Lingüística Hispánica*, 28, pp. 27-55.

De Benito Moreno, Carlota, Javier Pueyo & Inés Fernández-Ordoñez (2016): «Creating and designing a corpus of rural Spanish», in Nicoletta Calzolari *et al. Proceedings of the 13th Conference on Natural Language Processing* (KONVENS 2016), Ruhr-Universität Bochum, Bochum, pp. 78-83.

Domínguez Noya, Eva M.ª, Fco. Mario Barcala Rodríguez & Miguel Ángel Molinero Álvarez (2009): «Avaliación dun etiquetador automático estatístico para o galego actual: Xiada», *Cadernos de Lingua* 30-31, pp. 151-193.

Domínguez Noya, Eva M.ª (2013): *Etiquetaxe e desambiguación automáticas en galego: o sistema XIADA*. Tesis doctoral. Universidade de Santiago de Compostela. http://hdl.handle.net/10347/9587

Domínguez Noya, Eva M.ª (2016): «O etiquetador probabilístico de XIADA e o seu teito de acerto: a elaboración de regras lingüísticas», in Manuel González González (ed.): *Lingua, pobo e terra: estudos en homenaxe a Xesús Ferro Ruibal*. Santiago de Compostela: Xunta de Galicia - Centro Ramón Piñeiro para a investigación en humanidades, pp. 213-232.

García González, Marcos (2014): *Extracção de relações semânticas: recursos, ferramentas e estratégias*. Tesis doctoral. Universidade de Santiago de Compostela.

Graña Gil, Jorge (2000): *Técnicas de análisis sintáctico robusto para la etiquetación del lenguaje natural*. Tesis doctoral. Universidade da Coruña.

Guirao, José M.ª & Antonio Moreno-Sandoval (2006): «Morpho-syntactic tagging of the Spanish C-ORAL-ROM Corpus: methodology, tools and evaluation», in Yuji Kawaguchi, Susumu Zaima & Toshihiro Takagaki (eds.): *Spoken language corpus and linguistic informatics*. Amsterdam: John Benjamins, pp. 199-217. https://doi.org/10.1075/ubli.5.15mor

Manning, Cristopher. D. (2011): «Part-of-speech tagging from 97 % to 100 %: is it time for some linguistics?», in Alexander Gelbukh (ed.): *Computational linguistics and intelligent text processing, 12th International Conference, CICLing 2011, Proceedings*. Part I: *Lecture notes in computer science 6608*. Berlin: Springer, pp. 171-189. https://doi.org/10.1007/978-3-642-19400-9_14

Martín Zorraquino, M.ª Antonia & José Portolés Lázaro (1999): «Los marcadores del discurso», in Ignacio Bosque Muñoz & Violeta Demonte Barreto (eds.): *Gramática descriptiva de la lengua española*. Madrid: Espasa, pp. 4055-4143.

Martín Zorraquino, M.ª Antonia (2010): «Los marcadores del discurso y su morfología», in Óscar Loureda Lamas & Esperanza Acín Villa (coords.): *Los estudios sobre marcadores del discurso en español, hoy*. Madrid: Arco Libros, pp. 93-182.

Mendes, Amália, Raquel Amaro & M. Fernanda Bacelar do Nascimento (2004): «Morphological tagging of a spoken Portuguese corpus using available resources», in António Branco, Amália Mendes & Ricardo Ribeiro (eds.): *Language technology for Portuguese: shallow processing tools and resources*. Lisboa: Colibri [*s. p.*].

NGLE = Real Academia Española y Asociación de Academias de la Lengua Española (2009): *Nueva gramática de la lengua española*. Madrid: Espasa.

Pons Bordería, Salvador (2000): «Los conectores», in Antonio Briz Gómez (ed.): *¿Cómo se comenta un texto coloquial?* Barcelona: Ariel Practicum, pp. 193-220.

Portolés Lázaro, José (2001): *Marcadores del discurso*, Barcelona: Ariel Practicum.

Rojo Sánchez, Guillermo (2016): «*Citius, maius, melius*: Del CREA al CORPES XXI», in Johannes Kabatek (ed., con la colaboración de Carlota de Benito Moreno): *Lingüística de corpus y lingüística histórica iberorrománica*. Berlin: de Gruyter, pp. 197-212. https://doi.org/10.1515/9783110462357-010

EL *CORPUS DE APRENDICES DE ESPAÑOL* (CAES) Y SUS APLICACIONES PARA LA ENSEÑANZA/APRENDIZAJE DEL ESPAÑOL COMO LENGUA EXTRANJERA

The Corpus de Aprendices de Español *(CAES) and its applications to the teaching/learning of Spanish as a foreign language*

Ignacio Palacios Martínez (USC)
Francisco Mario Barcala Rodríguez (NLPgo Technologies S.L)
Guillermo Rojo (USC)

Resumen

El objetivo de este trabajo es presentar las características principales del CAES y sus aplicaciones en diversos ámbitos de la enseñanza y aprendizaje del español como lengua extranjera. En la primera parte se explica el origen de este proyecto así como las propiedades más importantes del corpus: diseño y estructura, proceso seguido para su compilación, tamaño, niveles y lenguas origen representadas. Se incluye aquí también una breve información referida a los procesos de etiquetación morfosintáctica y lematización, así como a las funcionalidades de la herramienta de consulta.

La segunda parte se centra en su totalidad en las aplicaciones y posibilidades de explotación del corpus que van desde investigaciones sobre las dificultades que tienen los aprendices en su aprendizaje y estudios contrastivos de interlengua hasta la elaboración de material de aula con datos extraídos del propio corpus y sus implicaciones para ámbitos como el diseño y desarrollo curricular, la formación del profesorado y la evaluación.

La aproximación adoptada es muy práctica de modo que las explicaciones teóricas van acompañadas de ejemplos ilustrativos que facilitan su comprensión.

Palabras clave: lingüística de corpus, corpus de aprendices, interlengua, español lengua extranjera.

Abstract

The objective of this paper is to present the main characteristics of CAES and its application to various areas of the teaching and learning of Spanish as a foreign language. The first section describes the origins of the project and its most salient features: design and structure, compilation process, size, learner-levels, and source languages represented. Also included here is some brief information on the processes of morphosyntactic tagging and lemmatisation, as well as a description of the functionalities of the query tool.

The second section focuses on its applications, including research into the difficulties that learners face during the language-learning process, contrastive studies of interlanguage, and the production of classroom material with data extracted from the corpus, as well as the implications here for areas such as curriculum design and development, teacher training, and evaluation.

The approach adopted is very practical, and theoretical explanations are accompanied by illustrative examples that allow for easy understanding by non-specialists.

Keywords: corpus linguistics, learner corpus, interlanguage, Spanish as a foreign language.

1. Introducción

El propósito de este trabajo reside en describir las características principales del *Corpus de aprendices de español* (CAES) (http://galvan.usc.es/caes) prestando atención especial a su explotación y aplicaciones por parte de docentes, investigadores y especialistas en el diseño y elaboración de materiales para la enseñanza y el aprendizaje del español como lengua extranjera. La aproximación utilizada es eminentemente práctica de modo que contribuya a despertar el interés por su consulta y anime a su utilización como herramienta de trabajo tanto para el estudio, la clase o la investigación.

Este capítulo está organizado en dos secciones o apartados principales; así, mientras que en la primera parte se explican el origen y desarrollo de este proyecto con sus características básicas más relevantes, la segunda se centra en las posibles aplicaciones de este corpus para la investigación del proceso de aprendizaje del español como lengua extranjera, la elaboración de actividades de aula, el diseño y desarrollo curricular, la formación del profesorado y la propia evaluación del trabajo del alumno. En lo posible, se ha intentado proporcionar ejemplos prácticos a modo de ilustraciones de todas estas aplicaciones.

2. Breve descripción del CAES

El *Corpus de aprendices de español* (CAES) es resultado del empeño personal de Francisco Moreno Fernández. En 2010, cuando fungía como director académico del Instituto Cervantes, nos encargó (a Guillermo Rojo e Ignacio Palacios) que preparásemos el diseño de un corpus de aprendices de español como lengua extranjera y de todos los procesos necesarios para desarrollarlo desde la recogida de muestras hasta su publicación. Presentamos el proyecto pocas semanas después y, tras los trámites administrativos oportunos, el Instituto Cervantes (IC) contrató con la Universidade de Santiago de Compostela (USC) el diseño, construcción, tratamiento lingüístico y explotación del CAES.

Analizadas las necesidades del IC para este proyecto, propusimos centrar el trabajo en estudiantes con seis L1 diferentes (árabe, chino mandarín, francés, inglés, portugués y ruso). Consideramos que, dada la implicación del IC en el proyecto, lo más adecuado era estructurar la recogida de pruebas en función de los niveles de conocimiento ya adquiridos por los participantes, con lo que quedaba claro desde el principio que no íbamos a tener estudiantes del nivel C2, de muy difícil o imposible localización. Por fin, decidimos montar el proceso de recogida mediante una aplicación informática que, en un entorno controlado (idealmente un centro de recursos del IC), pidiera a los participantes que cubrieran un formulario con sus datos y luego escribieran, en la misma sesión de trabajo, las pruebas correspondientes a su nivel de conocimientos. Al final de la sesión, la persona encargada del control de las pruebas las remitía a un servidor de la USC añadiendo un parte de incidencias producidas.

El procedimiento utilizado tiene dos grandes ventajas. La primera de ellas consiste en que no es necesario realizar la ardua y costosa tarea de escanear, transcribir y procesar textos manuscritos. La segunda procede del hecho de que el trabajo se ha simplificado también al ser posible utilizar el formato digital más adecuado en cada fase del procesamiento, como veremos más adelante. Con ello, la carga de trabajo de esta parte del proyecto resultó mucho más llevadera. Las desventajas radicaban en la necesidad de que las pruebas se realizaran en una sala con, por lo menos, conexión a Internet en la que los participantes pudieran trabajar en computadoras del centro o bien con sus propias máquinas. Aunque las ventajas superan con mucho a los inconvenientes, es claro que estos requisitos pueden suponer una dificultad adicional, puesto que no todos los centros del IC ni de las universidades colaboradoras disponen de instalaciones con estas características que puedan usar con facilidad.

Las pruebas solicitadas consistieron en la redacción de tres textos de entre 30 y 200 palabras cada uno en el caso de los estudiantes de los niveles (ya logrados, como hemos dicho) A1, A2 y B1, mientras que ese número se reducía a dos, pero con una extensión aproximada de 275-500 palabras, para los alumnos de los niveles B2 y C1. Consideramos adecuado fijar los temas de esas composiciones, así que se les pedía reservar una habitación en un hotel, hacer una reclamación de equipaje a una compañía aérea, contar una película, etc. Es decir, temas centrados en aspectos corrientes de la vida y planteables, por tanto, con distintos grados de conocimiento lingüístico. Muchos de estos temas suponen, como es lógico, una concentración de los elementos léxicos en ciertas zonas, lo cual provoca dificultades de importancia si se pretende proyectar los lemas obtenidos en el análisis de las producciones a cuestiones relacionadas con el análisis general del vocabulario adquirido por los estudiantes.

El diseño del proyecto incluía, naturalmente, la anotación morfosintáctica y la lematización de las producciones. Tras el análisis detenido de un conjunto de pruebas correspondientes a diferentes niveles y L1, llegamos a la conclusión de que los resultados del análisis automático iban a ser muy deficientes aunque dedicáramos tiempo y esfuerzo a la construcción de un corpus de entrenamiento, que debería tener muestras amplias por la multiplicidad de niveles y lenguas de origen. Optamos, en consecuencia, por utilizar una aplicación de anotación ya construida y concentrar el esfuerzo en la corrección manual (desambiguación) de sus resultados. FreeLing, de uso libre, nos pareció la mejor opción y, tras algunas pruebas iniciales, fue la que utilizamos finalmente, con un par de adiciones. Por una parte, Susana Sotelo de la USC preparó una rutina que combinaba la segmentación de oraciones con su *tokenización* y análisis con FreeLing. Para mayor comodidad en la desambiguación, permitimos que FreeLing seleccionara la etiqueta que presentaba los mejores resultados, pero añadiendo luego todas las etiquetas posibles para cada uno de los elementos resultantes del análisis automático. Por otro lado, María Paula Santalla del Río, también de la USC, construyó otra rutina para reconvertir algunas de las etiquetas usadas por la versión general de FreeLing a otras más conformes con el sistema de anotación usado habitualmente en nuestro grupo de la USC.

El paso siguiente consistió en el desarrollo de una aplicación que permitiera corregir los resultados automáticos haciendo las sustituciones, adiciones y eliminaciones tanto de etiquetas como de lemas. La tarea de corrección, dirigida por Paula Santalla, fue realizada por Alba Fernández Sanmartín y Marlén González González, lingüistas contratadas directamente en el pro-

yecto. Se hizo una fase de análisis por separado de los mismos textos, para comprobar el etiquetario, resolver dudas, fijar criterios y tomar decisiones en los casos dudosos.

El resultado de todo ello es un corpus de unos 600 000 elementos lingüísticos etiquetados y lematizados por lingüistas expertos en estas tareas. Se atribuyó etiqueta y lema también a aquellas formas inexistentes en español, de modo que *trayó*, por ejemplo, es etiquetada como la tercera persona de singular del pretérito de indicativo del verbo *traer*.

El paso final fue la construcción de una aplicación de consulta, desarrollada por Francisco Mario Barcala (NPLgo), que permitiera trabajar con todos los rasgos de los aprendices incluidos en la configuración del corpus (L1, nivel, sexo, edad, etc.) y añadir toda la potencia derivada de la anotación morfosintáctica y la lematización. Es posible, por ejemplo, recuperar todos los casos del verbo *ir* seguido del infinitivo de cualquier verbo a una distancia de uno o dos elementos (para que devuelva tanto casos del tipo *voy a salir* como *voy salir*). Como se ha indicado ya, esa búsqueda abstracta, basada en la consideración unitaria de todas las formas del paradigma del verbo *ir* y la detección de la forma de infinitivo de cualquier verbo, puede referirse a todos los estudiantes de nivel A1, a todos los estudiantes con portugués como L1 o a los estudiantes con nivel A1 y portugués como L1, que es el tipo de opción que proporciona habitualmente los datos necesarios para la investigación en este terreno. En los apartados 3.1 y 3.2 se muestran los resultados obtenidos en el análisis de algunos fenómenos como el mencionado.

Además de proporcionar la estadística simple, la ampliada y las concordancias, la aplicación de consulta permite acceder a toda la información correspondiente a la oración en la que se encuentra un determinado elemento. Como muestra la pantalla reproducida en la figura 1, se obtiene la secuencia de formas escritas originalmente, los elementos lingüísticos que la constituyen y los lemas a que pertenecen, con lo que se tiene acceso a la totalidad de los datos manejados, tanto en su forma original como con las capas de información añadida. Finalmente, todo ello es descargable en formato tsv (es decir, formato de texto con campos separados por tabuladores), que permite tanto su manejo directo como su integración en una hoja de cálculo o una base de datos.

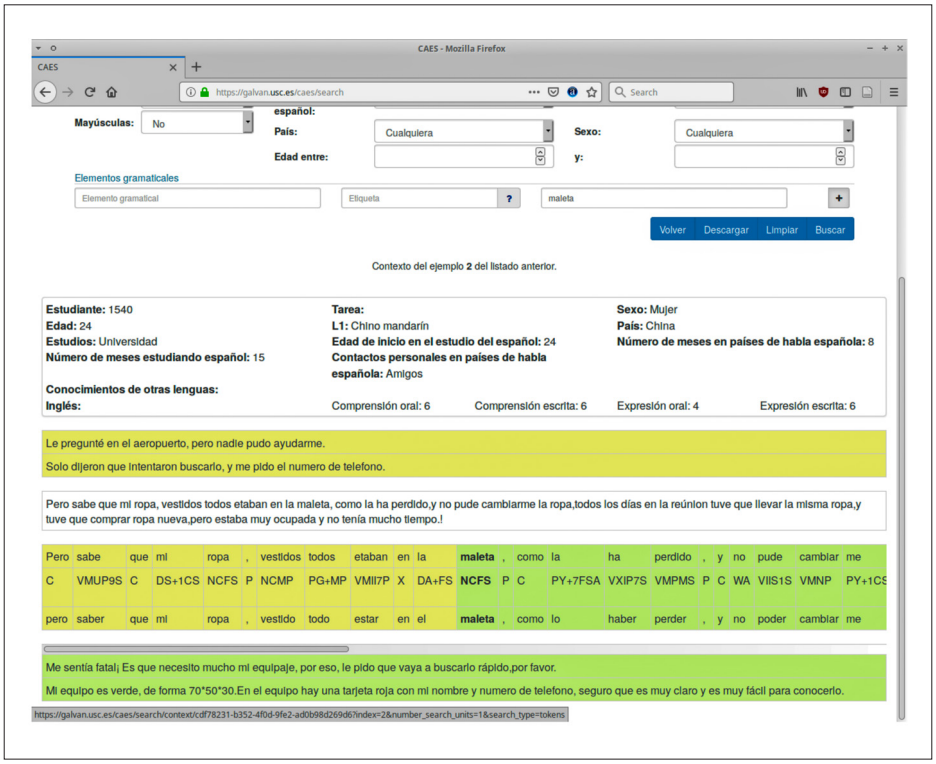


FIGURA 1. Ejemplo de los resultados obtenidos de una consulta simple en el corpus

En la figura 2 se muestra el flujo de procesamiento que han seguido las pruebas, desde que se organizaron las sesiones de recogida hasta que estas pasaron a formar parte de la aplicación de búsqueda. Como primer paso, los coordinadores habilitaban en la aplicación *online* de recogida diferentes sesiones para la realización de pruebas y los estudiantes realizaban las pruebas correspondientes que, al final de cada sesión, eran almacenadas en una base de datos relacional. Una vez eliminadas las muestras inválidas (duplicadas, incompletas o incluso vacías, entre otras causas), los textos sin las cabeceras fueron etiquetados con FreeLing y las etiquetas asignadas fueron convertidas en las equivalentes que se utilizaron en el proyecto, como ya hemos descrito anteriormente. A continuación, utilizando una herramienta desarrollada específicamente para esta tarea, se desambiguó manualmente el resultado de la etiquetación automática y, finalmente, las pruebas, ya etiquetadas y desambiguadas, se volvieron a unir a sus metadatos para generar un documento XML, lo que facilitó la realización de la carga de las pruebas en

una nueva base de datos específicamente diseñada para permitir realizar las búsquedas de la aplicación de consulta.

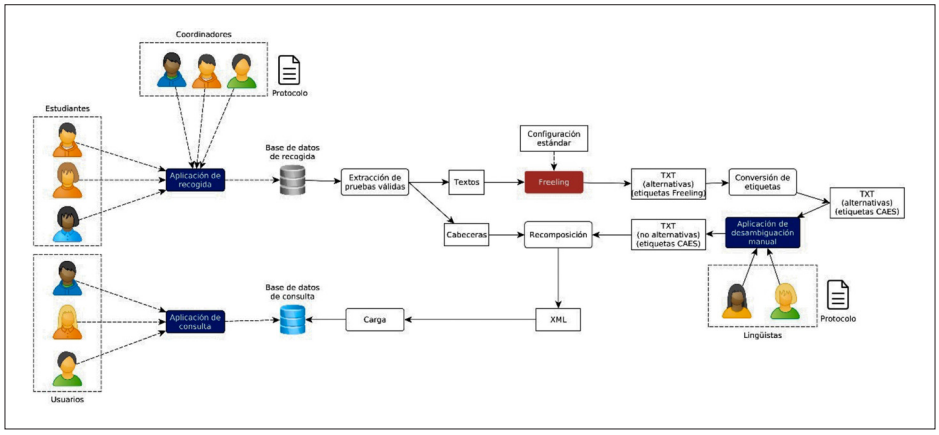


FIGURA 2. Flujo de procesamiento seguido por las pruebas desde el inicio hasta el final del proceso

3. APLICACIONES DEL CAES

Si bien existe un gran número de publicaciones en torno a las aplicaciones de la Lingüística de corpus y de los corpus de carácter general, por ejemplo, el *British National Corpus* (BNC) o el COBUILD en inglés, y el Corpus *de referencia del español actual* (CREA) o el *Corpus del español del siglo XXI* (CORPES XXI) en español, a la enseñanza de lenguas (Wichmann *et al.* 1997; Burnard & McEnery 2000; Kettemann & Marko 2000; Aston 2001; Granger *et al.* 2002; Hunston 2002; Aijmer 2009), no se puede afirmar lo mismo en lo que concierne a las aplicaciones concretas de los corpus de aprendices o de estudiantes a la didáctica de segundas lenguas. Esto se debe a dos razones fundamentales: la compilación de corpus de aprendices ha sido, hasta la fecha, mucho más restringida y reciente, y además estos corpus no representan la lengua de referencia o lengua meta, sino uno o varios tipos de interlengua, lo cual también conlleva ciertas limitaciones.

En nuestro trabajo partimos del planteamiento general de Leech (1997: 5) quien, al referirse a las posibles aplicaciones de los corpus lingüísticos a la enseñanza de lenguas, establece una división clara entre los usos directos e indirectos, entendiendo por los primeros «the direct use of corpora as resources for teaching», es decir, las aplicaciones de los corpus como recursos para la enseñanza y con un impacto concreto en la metodología de aula; en el caso

de los segundos, sin embargo, la contribución de los corpus a la didáctica de la lengua sería de carácter más secundario.

Adaptando este marco general a los corpus de aprendices en particular, que es lo que aquí nos ocupa, englobaríamos dentro de las aplicaciones directas las investigaciones realizadas sobre el aprendizaje del español/LE, la elaboración de actividades de aula y la confección de materiales didácticos. Como aplicaciones indirectas se encuadrarían aquellas que puedan tener una incidencia en la formación del profesorado, el diseño y desarrollo curricular, y en el ámbito de las pruebas lingüísticas y la evaluación. Nuestra exposición en las páginas que siguen estará organizada bajo estos epígrafes generales.

3.1 Investigación sobre el aprendizaje del español como lengua extranjera

Con cierta frecuencia, los profesores de lenguas nos mostramos más preocupados por el propio proceso de enseñanza que por el aprendizaje en sí, es decir, estamos más interesados en saber si una determinada técnica o actividad es realmente efectiva en el aula que por conocer detalles de cómo aprenden nuestros alumnos: qué dificultades se encuentran, cómo les gusta aprender, cuál es su motivación y su estilo de aprendizaje, qué estrategias ponen en marcha cuando no comprenden el significado de una palabra en un texto o no son capaces de descodificar un mensaje oral. Como resultado, no reparamos, a menudo, en que la información que podamos obtener sobre el proceso y experiencia de aprendizaje de nuestro alumnado podría ser de gran utilidad para iluminar, fortalecer e incluso mejorar determinados aspectos de nuestra docencia.

A tenor de esto, podemos distinguir tres grandes líneas de trabajo donde los corpus de aprendices, y el CAES en este caso, nos podrán proporcionar información de gran relevancia sobre el proceso de aprendizaje del español/LE de nuestro alumnado: (i) estudios comparativos entre el uso del español por parte de hablantes nativos y no nativos en la línea de lo que Granger (1998, 2002) definió como *Contrastive Interlanguage Analysis* (CIA), es decir, Análisis Contrastivo de Interlengua; (ii) investigaciones realizadas con el fin de confirmar o rechazar las hipótesis que apuntan a la existencia de una serie de áreas de la gramática española que presentan especial dificultad para el estudiante con el fin de aprovechar y revertir esta información en la propia enseñanza; (iii) estudios sobre la adquisición de morfemas gramaticales en el aprendizaje del español/LE en paralelo a los que se realizaron sobre el

español como L1 y sobre otras lenguas como el inglés en la que este tipo de investigaciones adquirieron un gran auge.

3.1.1 El análisis contrastivo de muestras de interlengua

La comparación de producciones escritas y/u orales de hablantes nativos de español con las de estudiantes de español/LE nos podría reportar datos de interés con el fin de dilucidar si hay determinadas categorías gramaticales, estructuras sintácticas, expresiones o términos léxicos, colocaciones, intensificadores, mitigadores y marcadores discursivos que son más o menos utilizados por los miembros de un grupo que por los del otro. Resulta interesante saber en qué medida el uso del español por parte de nuestros alumnos difiere del propio de los hablantes habituales para así, de ser necesario, introducir ajustes en nuestra enseñanza. Bajo el marco del Análisis Contrastivo de Interlengua sería posible, por ejemplo, contrastar muestras del CAES correspondientes al nivel C1 con otras del CORPES XXI de una tipología similar, teniendo en cuenta que los textos escritos que conforman este último están clasificados por tipo y género textual. Es evidente que los resultados de investigaciones de esta naturaleza serían limitados puesto que estos dos corpus no son totalmente comparables; para que así fuera, sería necesario compilar dos corpus *ad hoc*, uno de hablantes nativos y otro de aprendices, que siguieran criterios similares en cuanto a su estructura, método de compilación, tamaño, tipología textual, perfil de los participantes, etc., tal como ya existe para el inglés con el *International Corpus of Learner English* (ICLE) y el *Louvain Corpus of Native English Essays* (LOCNESS), que consisten en colecciones de ensayos de universitarios sobre temas semejantes[1]. Sin duda, esta sería una tarea pendiente para el futuro. No obstante, estudios basados en estos dos corpus con los que ya contamos, CAES y CORPES XXI, podrían apuntar al menos tendencias valiosas en una u otra dirección al contrastar estas producciones lingüísticas. Sin embargo, más factible y metodológicamente más justificable sería realizar comparaciones entre las muestras de participantes del CAES de distintas lenguas maternas en torno a una pregunta o hipótesis de investigación, habida cuenta de que en el momento presente están representadas en este corpus producciones escritas de alumnos de al menos seis lenguas maternas diferentes, tal como se explicó más arriba. Esto nos permitiría investigar en qué medida los estudiantes con distintas L1

[1] Para más información sobre estos corpus se puede consultar los enlaces siguientes: https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html https://www.learnercorpusassociation.org/resources/tools/locness-corpus/, último acceso 12 de marzo de 2019.

afrontan una misma cuestión de un modo similar o no. A modo de ejemplificación, veamos qué información podemos extraer del CAES cuando comparamos los ejemplos que aparecen con el verbo *gustar* en las muestras de estudiantes de francés y de inglés del mismo nivel, en este caso vamos a elegir el más bajo, es decir, A1. Seleccionamos para esta tarea muestras de estas dos lenguas por pertenecer a familias lingüísticas diferentes, lo que previsiblemente se concretará en resultados también diversos.

Para los estudiantes franceses de nivel A1, la herramienta de consulta recupera un total de 141 casos. Los datos de un primer análisis nos indican que para estos alumnos, a pesar de ser unos principiantes en el estudio del español, las construcciones con *gustar* no revisten, en principio, una gran dificultad. Los mayores problemas se derivan de la falta de concordancia entre sujeto y verbo (*no me gusta los deportes*; *no me gusta los persones que no só simpática*; *no me gusta películas de horror*) pero globalmente utilizan este verbo con corrección, atreviéndose incluso con estructuras un tanto complejas, teniendo en cuenta el grado de dominio del español por parte de estos alumnos como, por ejemplo, *a mi me gustaría trabajar en vuestra cruadilla*; *le gustaría ser un profesor de ciencias*. Si vemos ahora qué ocurre con los estudiantes de habla inglesa del mismo nivel, observamos que el número de ejemplos es más reducido que en el caso anterior, un total de 107, detectándose un mayor número de formas no estándares derivadas, no solo de la falta de concordancia como en el caso anterior (*me gusta todos los deportes*), sino también del uso pronominal (*Mi abrina se llama Ruolin, solo tiene 15 meses, se me gusta mucho*), orden de palabras (*me gusta arroz mucho*) y conjugación verbal (*Me gusto juego el chess; me gusto salir de noche a un restaurante*). Sería interesante profundizar más en esta cuestión tratando de identificar en una segunda fase las razones que pudieran justificar estos resultados, bien debido a factores de transferencia lingüística o de otra naturaleza. Para ello sería preciso examinar de modo detallado las muestras de los distintos niveles de los dos grupos de hablantes fijando una serie de parámetros de carácter gramatical, léxico e incluso pragmático que nos proporcionaría una información más veraz y fiable. Todo esto nos ayudaría a comprender si determinados procesos de adquisición del español son universales o, de lo contrario, están condicionados por la lengua origen de los sujetos.

3.1.2 Problemas en el aprendizaje

Los datos proporcionados por este corpus nos servirán para confirmar o rechazar la hipótesis de que los alumnos de español/LE tienden a tener difi-

cultades con el aprendizaje de determinadas áreas de la lengua a las que se suele prestar una atención especial en la mayor parte de las gramáticas pedagógicas, es decir, aquellos manuales gramaticales orientados de manera específica a la enseñanza de español/LE (Gómez, Pérez & Requeijo 1987; Benítez & Gelabert 1995; González Hermoso, Cuenot & Sánchez Alfaro 1999; Moreno 2001; Alonso Raya *et al.* 2005; Lieberman 2007; Areizaga Orube 2009; Palencia & Aragonés 2009; Romero Dueñas & González Hermoso 2011; Garnacho & Martín 2014; Villegas Galán & Blázquez Lozano 2014) o se identifican como tales en estudios globales de análisis de errores (Vázquez 1991; Santos Gargallo 1993, 2004). Entre estas áreas de especial dificultad destacan muchas que se pueden encuadrar dentro del nivel morfosintáctico de la lengua como son las diferencias entre los verbos *ser*, *estar* y *haber*, la conjugación y el uso del subjuntivo, los valores del indefinido frente al imperfecto, el uso de las preposiciones, en particular *para* frente a *por*, el género de los sustantivos, los artículos, la colocación de los pronombres, construcciones pasivas e impersonales con *se*, así como otras relacionadas más directamente con el léxico como los falsos amigos, expresiones idiomáticas o determinadas colocaciones, e incluso otras en las que el nivel léxico y el morfosintáctico interactúan, como ocurre con las diferencias entre *recordar* y *acordarse*, construcciones con el verbo *gustar*, etc. Todo esto lo debemos concebir dentro de un gran objetivo final que no será otro que conseguir que el alumno adquiera una competencia comunicativa con lo que todo ello implica (Consejo de Europa 2002)[2]. Lógicamente se aboga por un estudio de la gramática como medio para la obtención de la comunicación en todas sus facetas y vertientes en la línea que apuntan varios autores en el ámbito concreto de la enseñanza del español/LE (Matte Bon 1992a, 1992b, 1998, 2004; Gómez del Estal Villarino 2004; Lieberman 2007: 19; Areizaga Orube 2009: 2). Este es un principio general de partida que debe estar siempre presente en nuestra enseñanza.

[2] Para el Consejo de Europa (2002: 13) la competencia comunicativa engloba tres componentes: el lingüístico, el sociolingüístico y el pragmático. Cada uno de esos componentes incluye, a su vez, una serie de conocimientos, destrezas y habilidades. Los conocimientos están relacionados simplemente con el saber qué, mientras que las destrezas y habilidades con el saber cómo. Las competencias lingüísticas comprenden, por su parte, los conocimientos y las destrezas léxicas, fonológicas y sintácticas, así como otras dimensiones de la lengua como sistema. Las competencias sociolingüísticas tienen que ver con las condiciones socioculturales del uso de la lengua, mientras que las pragmáticas se refieren al uso funcional de los recursos lingüísticos sobre la base de contextos y escenarios de intercambios comunicativos. Este tipo de competencias engloban también todo lo relacionado con el dominio del discurso, la cohesión y la coherencia así como la identificación de los distintos tipos y formas de texto, la ironía y la parodia.

Sin profundizar ahora demasiado, sino a modo de esbozo y ejemplificación puesto que, tal como indicábamos más arriba, queremos animar al profesorado y especialistas en el uso del corpus, podemos ver en qué medida la confusión entre los verbos *ser* y *estar* resulta problemática para nuestros alumnos de español. Con este fin podemos llevar a cabo con la herramienta de consulta del CAES una búsqueda de la forma de primera persona del plural del presente del verbo *estar*, es decir, *estamos*. Una vez introducidos los datos en la aplicación del corpus, la herramienta de consulta recupera un total de 112 casos. Cuando analizamos estos resultados, observamos que en 20 de ellos, es decir, en alrededor del 18% del total, nos encontramos con un uso incorrecto y todos ellos corresponden, tal como por otra parte esperábamos, a muestras de estudiantes de niveles inferiores, A1, A2 e incluso B1, independientemente de su lengua origen, pues se presenta en alumnos cuya L1 es el árabe, el inglés o el chino, tal como muestran los ejemplos siguientes.

(1)

Nivel	L1	Ejemplo
A1	árabe	**Estamos** tres hermanos KHALID Y AHMED Y OFIANE.[3]
A2	árabe	nosotros **estamos** amigos de 18 años
A2	chino mandarín	Aquel entonces **estamos** felices cada día.
B1	inglés	**Estamos** amigas en Facebook y Tuenti también.

Así, la evidencia proporcionada por CAES confirma que efectivamente la diferencia entre *ser* y *estar* es problemática para los estudiantes de español y que, por lo tanto, el profesor debe prestarle una atención especial en el aula, realzando desde un punto de vista comunicativo cuándo se utiliza una forma u otra. Es evidente que el conocimiento de reglas de uso desde una perspectiva teórica no es suficiente puesto que es imprescindible saber llevar esas reglas a la práctica en la comunicación diaria. A tenor de esto y de los resultados anteriores, si profundizamos un poco más, observamos que cuando la forma investigada *estamos* va seguida de otra forma verbal, es decir, en construcciones durativas, los participantes del corpus no tienen en este caso mayores dificultades, como reflejan los siguientes ejemplos.

[3] Reproducimos los ejemplos seleccionados tal como se aparecen en el corpus sin introducir ningún cambio con respecto al original. En este caso los nombres propios aparecen todos ellos en mayúscula pues así los escribió el alumno. A este respecto es preciso tener en cuenta que en el CAES son los propios participantes los que introducen los datos de sus tareas escritas en la aplicación informática diseñada para tal propósito, evitándose de ese modo posibles errores en la transcripción derivados de falsas interpretaciones o de simples despistes del transcriptor.

(2)

Nivel	L1	Ejemplo
A1	árabe	Yo soy guitarrista en una banda de música y **estamos** jugando le música jazz.
A1	portugués	**Estamos** haciendo cordeiro.
A2	portugués	**Estamos** verificando algunos sitios.
B1	francés	**Estamos** yendo a Bruselas por el trabajo.

Además de lo anterior, sería totalmente factible e incluso sencillo trasladar todos estos ejemplos al aula y discutirlos con nuestros alumnos en la línea de lo que presentaremos en la sección siguiente (3.2). En este sentido, el corpus no solo nos proporciona información de interés sino que también nos puede dar algunas claves para llevar esto a la clase.

Pasando ahora a otra área de conflicto potencial, el uso de las preposiciones, sabemos también por experiencia, al igual que en el caso anterior, que suele ser una fuente de problemas. A la luz de este punto de partida, podemos explorar en qué medida los datos extraídos del CAES sustentan esta hipótesis, intentando averiguar también como segundo objetivo qué preposiciones en particular plantean mayores dificultades y en qué casos concretos. Utilizando de nuevo la herramienta de consulta del CAES, observamos que la preposición *de* es la más frecuente en el corpus con un total de 18 240 ejemplos, seguida de las preposiciones *en* (13 661 ejemplos) y *a* (9545 unidades). Por el contrario, *tras*, *según* y *hacia* son las menos comunes con unas cifras mucho más bajas, 13 en el primer caso y 27 en el de las dos últimas. Además, un primer análisis de estos resultados nos indica que las preposiciones *para* y *por* resultan ser las que entrañan mayor dificultad; *desde* ocuparía un punto intermedio. Profundizando un poco más en estos datos, detectamos que de los cuatro significados de *para* que reseña de Bruyne (1999: 678-681), a saber, finalidad, movimiento, período de tiempo y relación de personas, cosas y situaciones del mismo tipo, es con el tercero de ellos, es decir, el de valor durativo, con el que los estudiantes encuentran mayores dificultades al confundir con bastante frecuencia *para* con *por*. Los ejemplos siguientes tomados del corpus ilustran este hecho.

(3)

Nivel	L1	Ejemplo
A1	árabe	Pardon mi profesor, yo estoy en casa **para** tres días, porque yo soy muy enferma.
A2	portugués	¿Cuál es el precio total **para** 2 días?
A2	inglés	Yo vivi con mi papá y mi madrastra **para** dos meses.

Algo similar ocurre cuando *para* se utiliza para expresar dirección, movimiento.

(4)

Nivel	L1	Ejemplo
A1	portugués	Le gusta mucho su trabajo, pero también le gusta viajar **para** Alemania.
A2	portugués	Pero en Febrero, yo quiero viajar **para** Argentina.

Sin embargo, apreciamos también que con el significado de finalidad el número de confusiones es mucho menor. Esto es importante reseñarlo porque no solo resulta interesante centrarnos en las formas que no siguen la norma, sino también en los usos correctos, es decir, en aquello que hace bien el alumnado.

(5)

Nivel	L1	Ejemplo
A1	árabe	Yo leyó libos de gestión de empresas **para** amentar mi conocimiento.
A2	portugués	Voy a empezar mis estudios de derecho en la Universidad de Santiago de Compostela **para** ampliar mis conocimientos en la lengua española.

Si ahora llevamos a cabo un estudio similar con la preposición *por*, advertimos que los resultados del CAES nos confirman de nuevo nuestra pregunta de investigación puesto que los alumnos de todos los niveles tienen dificultades con su uso. Esta circunstancia puede derivarse de la multifuncionalidad de esta preposición; así, De Bruyne (1999: 681-690) reseña trece usos principales que van del valor agente en oraciones pasivas y duración a usos en oraciones concesivas (*por muy caro que sea*…) y exclamaciones (¡*por mis hijos que lo hago*!). Obsérvense los ejemplos siguientes:

(6)

Nivel	L1	Ejemplo
A1	árabe	El Mundo artístico es muy importante **por** mi.
A2	francés	Mi amigo Daniel tiene un problema con un pie y es un_poco dificile **por** el de caminar mucho.
B1	chino mandarín	¿Podrias contar me más detalles sobre Tianjin y reservar un alojamiento **por** mi?
C1	árabe	Fue seleccionada **por** el festival_ de_ cannes, pero no tenia premio.

Las muestras recuperadas también nos indican que existe una tendencia a confundir *por* con *durante*.

(7)

NIVEL	L1	EJEMPLO
A1	Inglés	Era un profesor **por** más de treinta años.
B1	Inglés	He viajando con tu compañía **por** diez años.[4]
B2	Árabe	Después de mi graduación viaje a España **por** dos semanas para asistir a un seminario de traducción Árabe.

Asimismo, se advierten confusiones cuando muchos de estos alumnos se refieren a medios de transporte, encontrándonos también con un número importante de colocaciones no gramaticales, tal como se puede apreciar en los ejemplos siguientes.

(8)

NIVEL	L1	EJEMPLO
A1	portugués	Soy aficionado **por** Real Madrid pero me gusta también el Barcelona.
A1	portugués	Soy responsable **por** todo el banco de dados de ventas de la compañía.
A2	inglés	Volveramos **por** autobús al centro de la cuidad.
A2	portugués	Mi familia y yo fuimos **por** carro y salimos muy temprano.
A2	chino mandarín	Está aficionado **por** películas y quieres ser un director en le futuro
B2	portugués	Tengo ganas **por** les conocer.

Al igual que en el caso anterior, es evidente que el profesor de español/LE tendrá que tener todo esto en cuenta en la planificación de sus clases y, sobre todo, a la hora de presentar y practicar estos contenidos gramaticales de la lengua.

3.1.3 Estudios sobre la adquisición de morfemas del español como lengua extranjera

En los años 70 y 80 del siglo pasado se realizaron, tomando como base el inglés como segunda lengua, una gran cantidad de investigaciones (Dulay & Burt 1973; Krashen & Terrel 1983; Krashen 1987) con el fin de averiguar si se podía hablar de un orden de adquisición similar o no al ya identificado para el inglés como L1, es decir, si en el proceso de aprendizaje de esta lengua

[4] En algunas variedades del español, este uso de *por* se considera totalmente gramatical.

se seguía una ruta semejante o no. Estos trabajos se basaban en el análisis de una serie de morfemas gramaticales tales como la *-s* de la tercera persona del presente simple, el uso de BE como cópula y en su valor progresivo, la formación del plural, el uso del genitivo con los nombres, los artículos, etc. La hipótesis que subyacía a estas investigaciones era que mediante el seguimiento longitudinal de un grupo de aprendices del inglés se podría llegar a la formulación de un orden de adquisición que supuestamente sería el mismo para todos los aprendices independientemente de su lengua materna. Esto reforzaría los presupuestos de la existencia de una Gramática Universal y pudiera tener ciertas implicaciones didácticas, ya que podría servir de base para el diseño curricular de programas de esta lengua (Pienemann 1989). Si bien estos estudios no estuvieron exentos de críticas (Hatch 1978; Ellis 2004), lo cierto es que efectivamente obtuvieron el éxito deseado puesto que fue posible llegar a establecer ese orden de adquisición para el inglés como L2 que, en líneas generales, resultó ser bastante semejante al orden de adquisición ya conocido e identificado previamente para el inglés como L1. Ya en fechas más recientes Housen (2002) realizó un estudio en el que investigaba la adquisición de morfemas verbales por parte de escolares norteamericanos de los cursos de tercero al undécimo, utilizando para ello muestras extraídas del *Corpus of Young Learner Interlanguage*, y, también en una línea semejante, Tono (2000) compara el orden de adquisición de morfemas gramaticales identificado a través de muestras japonesas de interlengua con el originalmente propuesto por Dulay & Burt (1973).

A la luz de estos resultados con referencia al inglés, Zobl & Liceras (1994) llevaron a cabo un trabajo en el que reconocen una secuencia de adquisición de morfemas en español/LE con cierto parecido a la del español nativo y en el que las marcas de presente indicativo, la negación con *no*, el presente durativo (*estoy hablando*), el futuro perifrástico (v*oy a hablar*) y el artículo indefinido (*un avión*) se adquieren antes, mientras que el uso de *ser* y *estar* como cópula, el pretérito perfecto y el imperfecto aparecen en estadios más avanzados del proceso (Baralo 1999). Estos resultados, en su conjunto, confirman los obtenidos por Van Naerssen (1980) en una investigación previa donde esta especialista también llegó a la conclusión de que los estudiantes americanos de español de su estudio mostraban tener pocos problemas a la hora de elegir el género del nombre cuando este aparecía modificado por un adjetivo. Sin embargo, la concordancia de género entrañaba para estos sujetos mayor complejidad que la concordancia de número. Estos dos resultados confirmaban, por otra parte, que no existía una gran diferencia entre el orden de adquisición del español como L1 y como L2. Caso contrario se evidencia

con respecto al tiempo verbal, más concretamente en lo que se refiere a la diferencia entre el pretérito e imperfecto, donde sí se observan diferencias importantes entre el español como lengua nativa y lengua segunda. Además, Van Naerssen (1980: 153), a tenor de su análisis, sugiere la existencia de un estadio en el desarrollo del español como L1 y como L2 en el que los hablantes tienden a utilizar una forma verbal básica a la que posteriormente le añaden una serie de flexiones. Esta forma básica pudiera ser fácilmente la tercera persona del singular del presente de indicativo o incluso el infinitivo, si bien es más probable que sea la primera que la segunda.

Dado que el CAES contiene muestras de alumnos cuyos niveles de competencia lingüística han sido muy controlados y que se pueden considerar fiables, sería posible también llevar a cabo trabajos similares a los anteriores, tomando como referencia los resultados previamente reseñados. Para ello sería aconsejable comenzar con el análisis de las muestras correspondientes a los alumnos de una lengua primera concreta para luego acometer estudios paralelos con las muestras de los alumnos de las distintas lenguas origen. Aun siendo conscientes de que el CAES no se puede considerar como un corpus longitudinal sino transversal ya que no se hace un seguimiento de la adquisición de los mismos aprendices a través del tiempo, el perfil académico de los participantes es bastante similar, lo que pudiera justificar un estudio de esta naturaleza, contribuyendo así a cubrir una de las lagunas existentes en la investigación de datos del español no nativo, tal como reconoce Liceras (1996: 239):

> Una de las grandes carencias con que nos enfrentamos a la hora de estudiar la gramática del español no nativo es la falta de estudios longitudinales que nos permitan analizar un corpus de datos suficiente.

3.2 Elaboración de actividades de aula con material del CAES

Si hasta el momento nos hemos venido refiriendo a aplicaciones del CAES centradas en el proceso de aprendizaje principalmente, en este apartado exponemos cómo se puede explotar este material de forma sencilla y clara en el aula. En esta línea partimos de que la aproximación pedagógica elegida estará centrada en torno al estudiante como principal protagonista del proceso enseñanza/aprendizaje (Nunan 1988a) y en el que el profesor adoptará el papel de guía, despertando la conciencia y sensibilidad gramatical y léxica del alumno en el uso del idioma en la línea de lo que Rutherford (1987) denomina «consciousness-raising» y «grammar awareness activities». Podemos

decir entonces que se trata de un aprendizaje por descubrimiento en el que la enseñanza de la lengua se hace de forma inductiva y los alumnos van progresando en su aprendizaje bajo la supervisión del profesor como si fueran pequeños investigadores de la lengua. Este tipo de acercamiento didáctico, que toma como recurso fundamental los datos proporcionados por un corpus, es lo que se ha venido en llamar en inglés *corpus/data-driven learning*, es decir, aprendizaje derivado de los datos de un corpus. Los trabajos en esta línea son numerosos (Tribble & Jones 1990; Johns 1991; Granger & Tribble 1998; Tribble 2015) y todos ellos inciden en los grandes beneficios que puede reportar la explotación de los corpus, ya sean generales o de aprendices, para la enseñanza de lenguas.

La primera actividad que proponemos, dirigida especialmente a estudiantes de los niveles iniciales aunque adaptable a otros más avanzados, se centra en el significado, uso y ortografía de la conjunción causal *porque,* tomando como base muestras del CAES producidas por alumnos del nivel A2. Nuestro propósito último es que el alumno sea capaz de expresar causalidad tanto oralmente como por escrito y lo sepa hacer con corrección y de acuerdo con el contexto. Hemos elegido este punto gramatical porque es un tema recurrente en los manuales de enseñanza del español, tiene que ver con varios ámbitos de la lengua (sintaxis, semántica, ortografía, discurso, etc.), circunstancia que lo hace especialmente interesante, y además hemos podido apreciar a través del análisis del material del corpus que plantea dificultades en su aprendizaje. En nuestra propuesta partimos de la presentación de datos para que el discente observe el funcionamiento de la lengua y vaya llegando a sus propias conclusiones. A medida que avanza la actividad, se incrementa también su grado de dificultad hacia un uso de la lengua más autónomo y creativo.

Para comenzar, presentamos a los alumnos los ejemplos siguientes para que los lean detenidamente.

1. Estoy muy bien porque ayer volví a mi casa de Hanolulu, Hawaii.
2. Es un hombre sincero muy generoso porque ayuda a la gente.
3. No me acuerdo del nombre y tampoco de la historia porque dormí todo el tiempo.
4. Le admiro porque él es honesto, divertido, responsable y un amigo muy bueno.
5. La admiro porque ella es simpática.

A continuación, les formulamos las preguntas siguientes:

1. ¿Qué palabra se repite en todas las oraciones anteriores? Subráyala, por favor.

2. Empareja los elementos de las 2 columnas de modo que tengan sentido.

Ejemplo: 1 e. Admiro a Mike **porque** es el hombre más simpático del mundo.

1. Admiro a Mike	a. porque es el verano aquí.
2. Hace mucho calor	b. porque es muy bonita y grande
3. Me gusta pasar tiempo con él	c. porque es un regalo de mi esposo.
4. Me ha encantado Barcelona	d. porque hizo mucho calor
5. La maleta es muy importante para mí	e. porque es el hombre más simpático del mundo.
Fuimos a la playa todos los días	f. porque es muy divertido.

3. ¿Qué significado expresa? Elige la respuesta correcta:

a. tiempo
b. causa
c. condición
d. lugar
e. fin

4. Ahora presta atención a su ortografía, es decir, a cómo se escribe ¿En qué casos se puede escribir como dos palabras separadas, es decir, *por qué*, como una palabra con tilde *porqué* y como una palabra sin tilde *porque*? Observa los ejemplos siguientes:

No entiendo *por qué* te enfadas.
¿*Por qué* no hiciste los deberes? *Porque* no tuve tiempo.
No saben el *porqué* de su comportamiento.
¿Cuáles son tus conclusiones?

.....

.....

5. Lee las oraciones siguientes e identifica los errores que encuentres. A continuación, escribe la oración correcta.

Ejemplo: Voy/llegar más tarde <u>en la</u> casa <u>por que</u> he quedado con amigos en un bar de tapas.
Voy a llegar más tarde a casa porque he quedado con amigos en un bar de tapas.

1. Voy llegar más tarde en la casa por que he quedado con amigos en un bar de tapas.

2. La visitaba y la preguntaba porque ha hecho esto.
3. No es bueno por la salud y tampoco por el bolsillo porque las cigaretas no son baratas.
4. Espero que tiene ya habitaciones libres por que es la alta período.
5. Le pregunte porque, pues havia lo suficiente por 40 personas.

6. Completa las oraciones siguientes utilizando una de las formas estudiadas:

1. No puedo opinar .....
2. ¿Por qué estás triste hoy? .....
3. Me cae bien Pedro .....
4. Prefiero callarme .....
5. ¿Por qué llegaste tarde? .....

La segunda tarea que proponemos dentro de esta sección está centrada en la enseñanza de la ortografía. Las muestras del CAES reflejan la dificultad que entrañan palabras con el grupo inicial *dif*, tales como «diferencia», «diferente», «difícil», «dificultad», ya que los alumnos tienden a escribirlas *diff*, posiblemente por transferencia del inglés. De hecho, nos encontramos con 38 casos de estas características. Ocurre algo similar con palabras con el grupo inicial *col* (*colaboración*, *colegio*, *colega*) que tienden a escribirlas con doble l, *collega*, *collaboración*, *collegio*, etc. Identificamos 15 ejemplos. Lo mismo pasa con palabras que en español contienen el grupo *cc* (*acción*, *construcción*, *dirección*, *ficción*, etc.) y que las suelen escribir con una *c* nada más.

Al igual que en la actividad anterior, les presentamos ejemplos donde se utilizan las palabras seleccionadas para que extraigan sus propias conclusiones y donde llamamos la atención sobre la ortografía de estos términos.

No noto mucho la di**f**erencia	Hablaba con di**f**icultad.
Me encanta cocinar di**f**erentes comidas.	La situación es di**f**ícil

Muchas gracias por tu co**l**aboración.
El co**l**egio de mis hijos es muy grande.
Le gusta mucho comer con sus co**l**egas de trabajo.

Esta es la dire**cc**ión de la casa.
Es un edificio de construc**cc**ión moderna.
La película es de ciencia fi**cc**ión.

A continuación, les proponemos la siguiente tarea:

diferente/diferir
dificultad
colegio
colega (el, la)
colaborar/colaborador

coleccionar/coleccionista
acción
construcción
ficción
elección

protección
sección
introducción
colección
reacción

Completa las oraciones siguientes con una de las palabras dadas en los recuadros superiores. Si tienes alguna duda sobre su significado, consulta el diccionario.

1. Mi amiga trabaja como ..... en el periódico local.
2. La ..... del presidente se pospone hasta mañana.
3. Tiene una ..... de monedas muy interesante.
4. La ..... de la nueva autopista se realizará en 2020.
5. La ..... de datos es importante hoy en día.
6. Siento ..... de tu opinión.
7. Debemos ponernos en .....

3.3 Elaboración de materiales didácticos

Los corpus de aprendices nos pueden proporcionar datos de interés que se pueden incorporar con cierta facilidad en gramáticas pedagógicas, diccionarios, glosarios y libros de texto concebidos específicamente para el aprendizaje del español/LE (véase apartado 3.1.2). Así, por ejemplo, en los diccionarios y glosarios, se podría llamar la atención a sus usuarios del término en español que están buscando con respecto a otra palabra paralela de su L1 que presenta una ortografía semejante pero que, sin embargo, posee un significado total o parcialmente diferente, lo que se conoce como «falsos amigos» (Chacón Beltrán 2006; Chamizo Domínguez 2008; Roca Varela 2015)[5]. De los datos del CAES se deriva que los estudiantes de español cuya lengua materna es el inglés tienden a identificar como semejantes, entre otros, los pares de palabras *suburb/suburbio*, *idiom/idioma*, *firm/compañía,* mientras que los de habla francesa tienen el mismo problema con *champagne/campiña*, *sentiment/impresión*, y los alumnos cuya L1 es el portugués con los pares *aula/clase*, *brincar/bromear*, *polvo/pulpo*, *romance/novela*; incluso hay términos que

[5] Véase un trabajo anterior (Rojo & Palacios 2016) donde analizábamos este tema con cierto detalle.

llevan a confusión a alumnos de distintas lenguas maternas, por ejemplo, *aplicar* con el sentido de solicitar o *largo* expresando gran tamaño. Además de esto, también se puede advertir al alumno sobre ciertas características peculiares del término en cuestión referidas a su naturaleza gramatical o a su uso. Así, a modo de ilustración, nuestras búsquedas en el CAES nos revelan que un número considerable de alumnos, sobre todo aquellos que tienen el portugués como L1, utilizan la combinación *ir* + infinitivo en lugar de la perífrasis *ir* + a + infinitivo, por ejemplo, *Julia fue comprar algo* en lugar de *Julia fue* ***a*** *comprar algo*; *fue conocer la Normandía* en lugar de *fue* ***a*** *conocer Normandía*. Detectamos también que existe una tendencia a designar los nombres de los países o continentes con el artículo determinado, y así se refieren a *la Libia*, *la Europa*, *la Francia*, *la Lituania*, *el Egipto* como en los ejemplos siguientes: *Después fue a la Francia*; *Va a volver a el Egipto*; *voy a ir a el España*; *mi país de nacionamento es la Colombia*; *prefiero el Marocco*. Lo mismo ocurre con la distinción entre *saber* y *conocer* (*Conoce español pero le gustaría seguir un curso de cocina*), las diferencias entre *qué* y *cuál* (*¿qué son las condiciones de admisión*?), cuestiones relacionadas con el número del sustantivo (*quiero saber más informaciones*), la colocación de los pronombres en la cláusula (*soy muy impaciente de conocer a vosotros*; *tengo ganas de conocer a vosotros*; *espero os conocer*). En realidad, sería posible añadir muchos más temas a este listado puesto que a medida que se va buceando en el corpus, más información de relevancia pedagógica se encuentra.

3.4 Formación del profesorado

Las aplicaciones de un corpus como el CAES a la formación del profesorado pueden ser de carácter general o más particular. Así, por ejemplo, se podría interpretar que algunas de las carencias detectadas en las muestras de los estudiantes en su uso del español pudieran tener su origen en debilidades o lagunas en la formación lingüística y pedagógica del profesorado. En ocasiones, cuando se analizan los programas de cursos y seminarios de formación docente de español/LE así como de otras lenguas, observamos que se suele poner el acento, entre otros, en teorías y aproximaciones didácticas, temas de innovación educativa y aplicaciones de las nuevas tecnologías sin abordar problemas concretos con los que el profesorado se tendrá que enfrentar en su día a día en el aula. En esta línea, sería posible entonces presentar a los docentes algunos de los ejemplos anteriores donde se advertían dificultades (*ser* frente a *estar*, preposiciones, uso del subjuntivo, colocación de los pronombres, falsos amigos, etc.) y tratar con ellos su tratamiento didác-

tico, incluyendo aquí actividades y ejercicios que se pudieran realizar con los alumnos en función de su nivel y de su perfil personal y académico. Asimismo, material de este tipo podría utilizarse muy fácilmente a la hora de abordar la corrección de errores y el tipo de comentarios que se deben emitir sobre la producción oral y escrita de los estudiantes, y sobre la calidad de sus trabajos. Sería incluso posible seleccionar muestras de los distintos niveles y solicitar al profesorado participante en un curso o seminario que clasificase esas muestras en diferentes niveles de dominio de la lengua (A1, A2, B1, etc.), tomando como base una serie de criterios definidos previamente. A todo lo anterior, habría que añadir que no estaría de más que se incluyese dentro de la formación del profesorado de lenguas extranjeras un pequeño módulo cuyo propósito sería la familiarización del profesorado con los corpus y su explotación didáctica, ya no solo como recurso y herramienta didáctica para sus clases sino también para su propio desarrollo profesional. Es evidente que a través de los corpus se pueden contrastar y aprender muchas cuestiones sobre el funcionamiento y uso de la lengua: frecuencias, colocaciones, registro, variación, actitudes de los hablantes, variables lingüísticas y extralingüísticas, etc.

3.5 Diseño y desarrollo curricular

Por regla general, los programas de los cursos de español/LE, al igual que ocurre con los programas de otras lenguas extranjeras, son diseñados por las autoridades educativas correspondientes, tomando como base los habituales principios y prácticas curriculares establecidos por especialistas en este campo (Nunan 1988b; Johnson 1989; Richards 2001; Medgyes & Nikolov 2010), así como documentos de referencia como el *Marco Común Europeo*, diccionarios y bibliografía especializada. Este es el caso del *Plan Curricular del Instituto Cervantes,* que fue elaborado por la Dirección Académica del Instituto Cervantes con la colaboración de un grupo de profesores y expertos, y los programas de las Escuelas Oficiales de Idiomas confeccionados por la administración educativa, es decir, el Ministerio de Educación y las Consejerías de Educación de las Comunidades Autónomas. Una vez que estos programas son publicados y puestos al servicio de la comunidad educativa, el profesorado elabora sus programaciones docentes tratando de adaptar los diseños curriculares oficiales a las especificidades y necesidades de su centro y de su alumnado. Los libros de texto también juegan un papel importante a este respecto porque, con frecuencia, condicionan las programaciones del profesorado pudiéndose convertir de facto en los verdaderos programas.

En función de lo anterior, el CAES podría utilizarse como material de referencia curricular pues contiene información real, ordenada y organizada por niveles que refleja el dominio del español, al menos en cuanto a su producción e interacción escrita, por parte de los alumnos que realizaron las tareas propuestas. Dado que la mayor parte de las muestras del corpus corresponden a alumnado de distintos centros del Instituto Cervantes de todo el mundo, sería factible investigar, por ejemplo, en qué medida existe una correspondencia entre la producción lingüística de estos alumnos con los niveles del *Marco Europeo* y con gran parte de los inventarios que se establecen en el *Plan Curricular del Instituto Cervantes* a propósito de la gramática, léxico, ortografía, técnicas y estrategias pragmáticas, etc. Es evidente que el proceso sería laborioso y que no sería posible cubrir todos y cada uno de los campos, pero sí serviría para proporcionar información necesaria para hacer un seguimiento de las actuales propuestas curriculares y tener así la base para la formulación de una serie de pautas de mejora. Este proceso parece pertinente pues se puede correr el peligro de que los diseños curriculares y, con ellos, los programas para cada uno de los niveles se queden en simples abstracciones incluyendo solamente aquello que se considera deseable pero que no resulta alcanzable en el día a día.

3.6 Otras posibles aplicaciones: evaluación, elaboración de pruebas

En los últimos años son más las voces (Mukherjee 2009; Barker 2010; Callies *et al.*, 2014; Callies & Götz 2015) que llaman la atención sobre la utilidad de los corpus de aprendices para el proceso de evaluación de la lengua extranjera. Así, por ejemplo, Callies & Götz (2015: 3) se refieren a esta cuestión en los términos siguientes:

> Generally speaking, learner corpora have the potential to increase transparency, consistency and comparability in the assessment of L2 proficiency, and in particular to inform, validate and advance the way L2 proficiency is assessed in the CEFR.

En este sentido el CAES, al recoger muestras representativas y contrastadas de cada nivel de competencia lingüística, nos proporciona información veraz sobre lo que son capaces de hacer los alumnos de cada nivel al mismo tiempo que nos revela sus limitaciones y dificultades. De este modo podría servir para aumentar la transparencia y solidez en la evaluación del español/LE. Sería factible también analizar las certificaciones existentes del español/LE y examinar en qué medida se adecuan a los diferentes grados de dominio

del alumnado, es decir, investigar hasta qué punto existe una correspondencia entre las actividades propuestas en estas pruebas y la competencia lingüística esperable de los alumnos de cada nivel. Además de esto, y entrando ya en el ámbito de la Lingüística computacional, material del CAES podría utilizarse también para el diseño e incluso la verificación de la efectividad de determinadas herramientas o instrumentos de detección de errores que de forma automática los clasifican de acuerdo con una serie de categorías fijadas de antemano (sintaxis, léxico, ortografía). Una experiencia de este tipo se ha realizado ya con dos corpus de aprendices de gallego con unos buenos resultados (Gamallo *et al.* 2015). En una línea similar, el corpus en parte o en su totalidad pudiera también constituir un buen recurso para la creación de una base de datos útil para la confección de pruebas y exámenes.

Referencias bibliográficas

Aijmer, Karen (ed.) (2009): *Corpora and language teaching*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.33

Alonso Raya, Rosario, Alejandro Castañeda Castro, Pablo Martínez Gila, Lourdes Miquel López, Jenaro Ortega Olivares & José Plácido Ruiz Campillo (2005): *Gramática básica del estudiante de español*. Barcelona: Difusión.

Areizaga Orube, Elisabet (2009): *Gramática para profesores de español como lengua extranjera*. Madrid: Ediciones Díaz de Santos.

Aston, Guy (ed.) (2001): *Learning with corpora*. Houston: Athelstan.

Baralo, Marta (1999): *La adquisición del español como lengua extranjera*. Madrid: Arco Libros.

Barker, Fiona (2010): «How can corpora be used in language testing?», in Anne O'Keeffe & Michael McCarthy (eds.): *The Routledge handbook of corpus linguistics*. New York: Routledge, pp. 633-645.

Benítez, Pedro & María José Gelabert (1995): *Breve gramática español lengua extranjera*. Barcelona: Difusión.

Burnard, Lou & Tony McEnery (eds.) (2000): *Rethinking language pedagogy from a corpus perspective*. New York: Peter Lang.

Callies, Marcus, María Belén Diez-Bedmar & Ekaterina Zaytseva (2014): «Using learner corpora for testing and assessing L2 proficiency», in Pascale Leclercq, Amanda Edmonds & Heather Hilton (eds.): *Measuring L2 proficiency: perspectives from SLA*. Clevedon: Multilingual Matters, pp. 71-90. https://doi.org/10.21832/9781783092291-007

Callies, Marcus & Sandra Götz (eds.) (2015): *Learner corpora in language testing and assessment*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.70

Chacón Beltrán, Rubén (2006): «Towards a typological classification of false friends (Spanish-English)», *Revista Española de Lingüística Aplicada* 19, pp. 29-39.

Chamizo Domínguez, Pedro (2008): *Semantics and pragmatics of false friends*. New York: Routledge.

Consejo de Europa (2002): *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*. Madrid: Ministerio de Educación, Cultura y Deporte.

De Bruyne, Jacques (1999): «Las preposiciones», in Ignacio Bosque & Violeta Demonte (eds.): *Gramática descriptiva de la lengua española*. Madrid: Espasa, pp. 657-704.

Dulay, Heidi C. & Marina K. Burt (1973): «Should we teach children syntax?», *Language Learning* 23, pp. 245-258. https://doi.org/10.1111/j.1467-1770.1973.tb00659.x

Ellis, Rod (2004): *Understanding second language acquisition*. Oxford: Oxford University Press.

Gamallo, Pablo, Marcos García, Iria del Río & Isaac González (2015): «Natural language processing for automatic error detection», in Marcus Callies & Sandra Götz (eds.): *Learner corpora in language testing and assessment*. Amsterdam: John Benjamins, pp. 35-57.

Garnacho López, Pilar & María Dolores Martín Acosta (2014): *Diccionario de dudas del estudiante de español como lengua extranjera*. Madrid: SGEL.

Gómez, María Dolores, María Jesús Pérez & María H. Requeijo (1987): *Gramática española para extranjeros*. Santiago de Compostela. Servicio de Publicaciones de la Universidad.

Gómez del Estal Villarino, Mario (2004): «Los contenidos lingüísticos o gramaticales. La reflexión sobre la lengua en el aula de E/LE: criterios pedagógicos, lingüísticos y psicolingüísticos», in Jesús Sánchez Lobato & Isabel Santos Gargallo (dirs.): *Vademécum para la formación de profesores: enseñar español como lengua segunda (L2) / lengua extranjera (LE)*. Madrid: SGEL, pp. 767-787.

González Hermoso, Alfredo, Jean-Rémy Cuenot & María Sánchez Alfaro (1999): *Gramática del español lengua extranjera*. Madrid: Edelsa.

Granger, Sylviane (ed.) (1998): *Learner English on computer*. Longman: New York.

Granger, Sylviane (2002): «A bird's eye view of learner corpus research», in Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson (eds.): *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins, pp. 3-33. https://doi.org/10.1075/lllt.6.04gra

Granger, Sylviane, Joseph Hung & Stephanie Petch-Tyson (eds.) (2002): *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.6

Granger, Sylviane & Chris Tribble (1998): «Exploiting learner corpus data in the classroom: form-focused instruction and data-driven learning», in Sylviane Granger (ed.): *Learner English on computer*. New York: Longman, pp. 199-209.

Hatch, Evelyn M. (ed.) (1978): *Second language acquisition: a book of readings*. Rowley, Mass: Newbury House Publishers.

Housen, Alex (2002): «A corpus-based study of the L2-acquisition of the English verb system», in Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson (eds.): *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins, pp. 77-116.

Hunston, Susan (2002): *Corpora in applied linguistics*. Cambridge: Cambridge University Press. https://doi.org/10.1075/lllt.6.08hou

Instituto Cervantes. *Plan curricular*. https://cvc.cervantes.es/ENSENANZA/biblioteca_ele/plan_curricular/default.htm [última consulta: 12/03/2019]

Johns, Tim (1991): «Should you be persuaded: two examples of data-driven learning», *English Language Research Journal* 4, pp. 1-16.

Johnson, Robert K. (1989): *The second language curriculum*. Cambridge: Cambridge University Press.

Kettermann, Bernhard & Georg Marko (eds.) (2000): *Teaching and learning by doing corpus analysis: proceedings of the Fourth International Conference on Teaching and Language Corpora*. New York: Rodopi.

Krashen, Stephen (1987): *Principles and practice in second language acquisition*. Englewood Cliffs: Prentice Hall.

Krashen, Stephen & Tracy D. Terrel (1983): *The natural approach: language acquisition in the classroom*. Oxford: Pergamon.

Leech, Geoffrey (1997): «Teaching and language corpora: a convergence», in Anne Wichmann, Steven Fligelstone, Tony McEnery & Gerry Knowles (eds.): *Teaching and language corpora*. London: Longman, pp. 1-23.

Liceras, Juana M. (1996): *La adquisición de las lenguas segundas y la gramática universal*. Madrid: Síntesis.

Lieberman, Dorotea I. (2007): *Temas de gramática del español como lengua extranjera*: *una aproximación pedagógica*. Buenos Aires: Editorial Universitaria de Buenos Aires.

Matte Bon, Francisco (1992a): *Gramática comunicativa del español. De la lengua la idea, I*. Barcelona: Difusión.

Matte Bon, Francisco (1992b): *Gramática comunicativa del español. De la idea a la lengua, II*. Barcelona: Difusión.

Matte Bon, Francisco (1998): «Gramática, pragmática y enseñanza comunicativa del español», *Carabela* 43, pp. 53-79.

Matte Bon, Francisco (2004): «Los contenidos funcionales y comunicativos», in Jesús Sánchez Lobato & Isabel Santos Gargallo (dirs.): *Vademécum para la formación de profesores: enseñar español como lengua segunda (L2) / lengua extranjera (LE)*. Madrid: SGEL, pp. 811-834.

Medgyes, Péter & Marianne Nikolov (2010): «Curriculum development in foreign language education: the interface between political and professional education», in Robert B. Kaplan (ed.): *The Oxford handbook of applied linguistics*. Oxford: Oxford University Press, pp. 264-274. https://doi.org/10.1093/oxfordhb/9780195384253.013.0018

Moreno, Concha (2001): *Temas de gramática*. Madrid: SGEL.

Mukherjee, Joybrato (2009): «The grammar of conversation in advanced spoken learner English: learner corpus data and language-pedagogical implications», in Karen Aijmer (ed.): *Corpora and language teaching*. Amsterdam: John Benjamins, pp. 203-230. https://doi.org/10.1075/scl.33.17muk

Nunan, David (1988a): *The learner-centred curriculum*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524506

Nunan, David (1988b): *Syllabus design*. Oxford: Oxford University Press.

Palencia, Ramón & Luis Aragonés (2009): *Gramática y uso del español para extranjeros: teoría y práctica* [distintos niveles]. Madrid: S.M.

Pienemann, Michael (1989): «Is language teachable? Psycholinguistic experiments and hypothesis», *Applied Linguistics* 10, pp. 52-79. https://doi.org/10.1093/applin/10.1.52

Richards, Jack C. (2001): *Curriculum development in language teaching*. Cambridge: Cambridge University Press.

Roca Varela, María Luisa (2015): *False friends in learner corpora: a corpus-based study of English false friends in the written and spoken production of Spanish learners*. Bern: Peter Lang. https://doi.org/10.3726/978-3-0351-0841-5

Rojo, Guillermo & Ignacio M. Palacios (2016): «Learner Spanish on computer: The CAES 'Corpus de Aprendices de Español' project», in Margarita

Alonso-Ramos (ed.): *Spanish learner corpus research: current trends and future perspectives*. Amsterdam: John Benjamins, pp. 55-86. https://doi.org/10.1075/scl.78.03roj

Romero Dueñas, Carlos & Alfredo González Hermoso (2011): *Gramática del español lengua extranjera*. Madrid: Edelsa.

Rutherford, William (1987): *Second language grammar: learning and teaching*. London: Routledge.

Santos Gargallo, Isabel (1993): *Análisis contrastivo, análisis de errores e interlengua en el marco de la lingüística contrastiva*. Madrid: Síntesis.

Santos Gargallo, Isabel (2004): «El análisis de errores en la interlengua del hablante no nativo», in Jesús Sánchez Lobato & Isabel Santos Gargallo (dirs.): *Vademécum para la formación de profesores: enseñar español como lengua segunda (L2) / lengua extranjera (LE)*. Madrid: SGEL, pp. 391-410.

Tono, Yukio (2000): «A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes», in Lou Burnard & Tony McEnery (eds.): *Rethinking language pedagogy from a corpus perspective*. New York: Peter Lang, pp. 123-132.

Tribble, Chris (2015): «Teaching and language corpora: Perspectives from a pedagogical journey», in Agnieszka Leńko-Szymańska and Alex Boulton (eds.): *Multiple affordances of language corpora for data-driven learning*. Amsterdam: John Benjamins, pp. 37-62. https://doi.org/10.1075/scl.69.03tri

Tribble, Chris & Glyn Jones (1990): *Concordances in the classroom*. Harlow: Longman.

Van Naerssen, Margaret (1980): «How similar are Spanish as a first language and Spanish as a foreign language?», in Robin C. Scarcella & Stephen D. Krashen (eds.): *Research in second language acquisition: selected papers of the Los Angeles Second Language Acquisition Research Forum*. Rowley, Mass: Newbury House Publishers, pp. 146-154.

Vázquez, Graciela E. (1991): *Análisis de errores y aprendizaje de español/lengua extranjera*. Frankfurt am Main: Peter Lang.

Villegas Galán, María de los Angeles & María Jesús Blázquez Lozano (2014): *Universo gramatical: gramática de referencia para estudiantes de español*. Madrid: Edinumen.

Wichmann, Anne, Steve Fligelstone, Tony McEnery & Gerry Knowles (1997): *Teaching and language corpora*. London: Longman.

Zobl, Helmut & Juana Liceras (1994): «Functional categories and acquisition orders,» *Language Learning* 44/1, pp. 159-180. https://doi.org/10.1111/j.1467-1770.1994.tb01452.x

MULTIFUNCIONALIDAD DE LOS CORPUS PARALELOS, EJEMPLIFICADA CON EL CORPUS ALEMÁN / ESPAÑOL PaGeS

Multifunctionality of parallel corpora, exemplified by German-Spanish corpus PaGeS

Irene Doval, Tomás Jiménez
Universidade de Santiago de Compostela

Resumen

En este capítulo se muestra cómo un corpus paralelo, ejemplificado con el corpus bilingüe alemán / español PaGeS, puede convertirse en una herramienta multifuncional apta para muy diferentes tipos de usuarios, si su composición, recuperación de información, posibilidades de búsqueda y visualización cumplen ciertos criterios. El corpus PaGeS consta de dos partes, un núcleo y unos suplementos. En este trabajo se describen los diferentes pasos seguidos en la construcción del corpus nuclear, una colección de narrativa contemporánea española y alemana. Esta descripción incluye la preparación manual de los textos, el alineado automático y la revisión manual. Se explica el acceso y visualización de los datos, así como los diferentes niveles de búsqueda adaptados a distintos usuarios. Finalmente se hace un balance y se esbozan las líneas de desarrollo futuro.

Palabras clave: lingüística de corpus, corpus paralelo, alineado del corpus, visualización del corpus

Abstract

This chapter shows how a parallel corpus, exemplified by the German-Spanish bilingual corpus PaGeS, can become a multifunctional tool, suitable for many different types of users, provided that its composition, information retrieval and the possibilities of searching and visualization meet certain criteria. The PaGeS corpus consists of two parts,

the core corpus and the supplements. In the present study we describe the different steps taken in the construction of the core corpus, which consists of a collection of contemporary Spanish and German narrative texts and their translations. This description includes the manual preparation process of the texts and the manual and automatic procedure of sentence alignment. It explains the access and the visualization of the data, as well as different search levels according to the needs of different users. The study ends with an evaluation and outline of the next steps to be taken in the future.

Keywords: corpus linguistics, parallel corpora, corpus alignment, corpus visualization

1. Introducción: Los corpus paralelos

Desde los años 90 las investigaciones lingüísticas basadas en corpus se han generalizado como el método estándar, revolucionando todas las disciplinas lingüísticas. Los primeros corpus se remontan ya a la década de 1960, eran exclusivamente monolingües y mayoritariamente de textos ingleses. Durante más de dos décadas, los corpus continuaron siendo monolingües, hasta que la creación del Corpus Paralelo Inglés/Noruego[1] (ENPC) (Johansson & Hofland 1994) y su proyecto hermano Corpus Paralelo Inglés/Sueco (Aijmer & Altenberg 1996: 79 ss.), a comienzos de los años 90, marcaron el comienzo de la era de los corpus paralelos y fueron determinantes en su ulterior evolución. Siguiendo este modelo se compilaron varios corpus en los años siguientes (Hasselgard 2015: 4). Este rápido desarrollo de los corpus paralelos le permitió a Lars Borin, ya en 2002 afirmar que «in the last decade or so, parallel corpus linguistics has emerged as a distinct field of research within corpus linguistics, itself a fairly young discipline» (Borin 2002: 1). Desde entonces se sucedieron conferencias y workshops dedicados exclusivamente a los corpus paralelos y comparables[2], a su creación, compilación, anotación y procesado. Asimismo, una amplia variedad de corpus paralelos fueron creados para diferentes lenguas y con diferentes objetivos.

Este es el caso de los recursos derivados de textos producidos por las diferentes instituciones de la Unión Europea (*vid.* Steinberger *et al.* 2014 para una visión general). Estos están incluidos, por ejemplo, en el corpus Multilin-

[1] Hubo corpus paralelos anteriores, como el de Filipovic o el *Canadian Hansard Corpus*, pero permanecieron como iniciativas aisladas sin continuidad. Para una visión más detallada sobre la evolución de los corpus paralelos, *vid.* Doval & Sánchez (2019: 1-19).

[2] Según la terminología comúnmente aceptada (McEnery & Hardie 2012: 20), en los corpus multilingües se distingue entre corpus comparables y paralelos. A diferencia de los corpus paralelos, los textos de un corpus comparable no son traducciones unos de otros, sino que son textos monolingües en diferentes idiomas que comparten tema, tipología textual y registro con un origen y un alcance similares.

gwis (Clematide *et al.* 2016), una herramienta de búsqueda desarrollada en la Universidad de Zurich que cubre los debates del Parlamento Europeo en siete lenguas, incluyendo alemán y español. Forman parte también de la gran colección de corpus paralelos integrados en el proyecto Opus de Jörg Tiedemann (Tiedemann 2012). Además de documentos administrativos de las instituciones europeas, Opus contiene también textos periodísticos y algunas colecciones menores de diferentes fuentes en línea, como subtítulos y documentación técnica. Otro corpus paralelo multilingüe es InterCorp, compilado en la Universidad Carolina de Praga (Čermák 2019). Abarca 40 lenguas, entre ellas el alemán y español. Además de los textos de la UE y subtítulos, incluye también textos de ficción, utilizando el checo como la lengua pivote para el alineado.

Por último, hay que mencionar un recurso que tiene un amplísimo uso, Linguee, que es una herramienta en línea que combina ciertas prestaciones de diccionarios bilingües con ejemplos de uso alineados paralelamente, acercándose también a lo que es una memoria de traducción. Actualmente cubre 25 lenguas, según informaciones propias y, aunque dispone de una cierta variedad de textos, la mayoría están vinculados al tipo de texto administrativo o comercial de la Unión Europea.

El objetivo de este capítulo es presentar el corpus paralelo alemán / español, PaGeS[3], mediante una descripción de su contenido y de sus características distintivas (§ 2); la sección 3 está destinada a explicar el procesado, segmentación y alineado de los textos, mientras que la indexación, posibilidades de búsqueda y visualización de los datos están comentados en la sección 4. Por último, en el § 5 se hace un balance en el que se incluyen las funcionalidades previstas no implementadas, así como potenciales ampliaciones futuras.

2. El corpus paralelo alemán/español PaGeS: composición

Como se ha indicado en el §1, la mayoría de los recursos multilingües disponibles para el par de lenguas alemán/español se limitan a variedades textuales específicas, principalmente lengua jurídica, administrativa o técnica. Además, en la inmensa mayoría de los casos no se puede determinar con precisión ni la lengua original, ni el proceso de traducción llevado a cabo. Así que, tanto por el tipo de textos como por las características de su producción, estos recursos, aunque indudablemente valiosos, adolecen de limitaciones como

[3] Este corpus que se encuentra disponible en línea desde 2016 (www.corpuspages.eu) está siendo elaborado por el equipo de investigación SpatiAlEs, de la Universidad de Santiago de Compostela, dirigido por Irene Doval. Este grupo consta de miembros de las universidades de Santiago de Compostela, Salamanca, Complutense de Madrid y Valladolid. El proyecto de elaboración del corpus está siendo financiado por el Ministerio de Economía y Competitividad (FFI2013-42571-P y FFI2017-85938-R).

herramientas multifuncionales; en efecto, los corpus precedentes presentan ciertas desventajas tanto para la investigación en lingüística contrastiva y traducción como para la enseñanza y aprendizaje de lenguas extranjeras, tareas fundamentales de nuestro grupo de investigación (Doval 2017b: 128).

La lengua comercial, administrativa y jurídica presenta poca variedad léxica y morfosintáctica en las estructuras gramaticales más empleadas, ya que se tiende al uso de fórmulas estereotipadas. Asimismo, este tipo de discurso no resulta adecuado para la enseñanza de la lengua con fines no específicos, por presentar un nivel de dificultad demasiado elevado.

Para la investigación en lingüística contrastiva y para el estudio de fenómenos relacionados con la lengua traducida, es indispensable identificar inequívocamente la lengua original y la lengua traducida, esto es, determinar cuál es la lengua fuente y cuál la lengua meta. Además, resulta también fundamental conocer el proceso de traducción, esto es, si se trata de una traducción directa entre el alemán y el español o, si por el contrario, se trata de una traducción indirecta a través de una tercera lengua.

Por último, hay que señalar que para cualquier investigación lingüística o recurso didáctico es condición indispensable la calidad de los materiales que constituyen la base empírica del trabajo. Los recursos mencionados, a excepción de los textos de la Unión Europea, no proporcionan estándares de calidad ni para los textos originales ni para las traducciones, ya que no han sido sometidos a controles de calidad comprobables.

Estas limitaciones de los recursos existentes han constituido la motivación inmediata para la creación del corpus PaGeS. Dado que las actividades del grupo de investigación se centran en la lingüística contrastiva y en la enseñanza de lenguas, resulta imprescindible asegurar la calidad del material, tanto con respecto a los textos originales como a su traducción. Para garantizarla, los textos tienen que haber pasado necesariamente algún control de calidad. Por otro lado, hay que tener en cuenta que la compilación de un corpus paralelo, frente a uno monolingüe, tiene enormes limitaciones en cuanto al material disponible, ya que la inmensa mayoría de los textos no se traducen y, cuando lo están, solo una pequeña parte pasa algún tipo de control. Por ello, la única vía disponible es acudir a materiales publicados por editoriales reconocidas, donde originales y traducciones hayan sido sometidos a un estricto control de calidad (Doval 2017b, 2018).

Por estos motivos, en PaGeS se ha compilado un corpus nuclear de textos narrativos escritos después de 1960, aunque con un claro predominio de obras a partir del año 2000. En cuanto al género, están integrados tanto textos de ficción como una pequeña parte de no ficción, ya que constituyen

la mayor parte de los recursos bilingües disponibles. Actualmente (abril de 2019) están incluidos textos de 140 obras originales (en español o alemán, además de una pequeña parte en una tercera lengua, *vid.* figura 1) y sus traducciones, con un tamaño total de más de 25 millones de palabras, unos 28 millones de tokens y 891.265 bisegmentos, es decir, pares de unidades alineadas (oraciones o segmentos más pequeños). En las obras literarias se ha tratado de atender a la variedad de géneros, con cierto predominio de la literatura infantil, así como a la variedad dialectal. En los textos españoles hay una amplia muestra de autores americanos y, en los textos alemanes, los autores suizos y austríacos están también representados. Entre las obras de no ficción se encuentran ensayos políticos e históricos, así como prosa de divulgación científica. Por otro lado, se ha velado por el equilibrio en las direcciones de la traducción, por lo que el corpus ofrece una composición proporcionada en cuanto a la lengua original, tal como se recoge en la figura 1. La tabla 1 muestra la composición del corpus nuclear, disponible ya en línea, con la indicación del número de obras y de palabras distribuidas según la lengua original.

Lengua	Obras	Palabras	Bisegmentos
Alemán Original	66	5 354 413	406 029
Español Traducción < Alemán	66	5 601 475	406 029
Español Original	56	5 221 244	332 955
Alemán Traducción < Español	56	5 145 174	332 955
Alemán Traducción < 3ª Lengua	18	2 101 211	152 347
Español Traducción < 3ª Lengua	18	2 157 039	152 347
Total	140 (x 2)	25 580 556	891 265 (x2)

TABLA 1. Composición del corpus nuclear de PaGeS (abril de 2019)

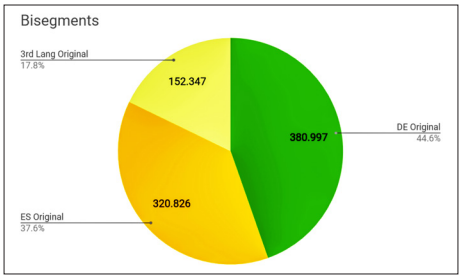


FIGURA 1. Distribución de los bisegmentos por lengua original (abril 2019)

Además de este material nuclear, el corpus PaGeS ofrece otros materiales complementarios, que no reúnen todos los requisitos enumerados anteriormente, pero que, para determinados fines, pueden suponer un complemento valioso. En el momento actual estos materiales están compuestos por las versiones alemana y española del corpus Europarl y de los discursos TED.

El Europarl (Release v7) está constituido por las actas del Parlamento Europeo de los años 1996 a 2011. El corpus ha sido extraído y alineado automáticamente en corpus paralelos con la versión inglesa por Koehn (Koehn 2005). La versión utilizada en PaGeS para el par alemán/español procede de Opus (Tiedemann 2012). La lengua original no está consignada en la mayoría de las ocasiones y suponemos que, conforme a la praxis habitual de las instituciones europeas, se ha partido de la versión inglesa para la traducción a las respectivas lenguas, por lo que una traducción directa entre español y alemán constituye un hecho excepcional.

Los textos de TED proceden de las traducciones al alemán y al español de las transcripciones de las charlas TED, todas ellas originalmente pronunciadas en inglés. Las traducciones han sido llevadas a cabo por traductores voluntarios y sometidas a un proceso de revisión. Usamos la versión en xml ofrecida por Web Inventory of Transcribed and Translated Talks (WIT3) (Cettolo /Girardi /Federico 2012). Posteriormente han sido alineadas, usando el LF-Aligner y sometidas a diversos procesos de limpieza automática antes de su indexación. La tabla 2 presenta las estadísticas correspondientes a estos suplementos.

	Europarl		**TED**	
	Palabras	**Bisegmentos**	**Palabras**	**Bisegmentos**
Alemán	44 303 154	1 882 959	4 061 184	234 328
Español	49 860 899		4 248 416	

Tabla 2. Europarl y TED integrado en PaGeS

3. Procesado, segmentación y alineado de los textos

Esta sección se refiere exclusivamente al corpus nuclear, esto es, la parte del corpus PaGeS que procede de obras narrativas publicadas en editoriales. Después de la selección de textos, estos han de ser preparados para el alineado. Para ello ambas versiones se almacenan como archivos txt con la codificación común UTF-8. En una primera fase se trata de reducir lo máximo posible el ruido y de revisar los textos para que el texto fuente original y el meta sean lo

más paralelos posibles, a fin de conseguir unos resultados más precisos. De ahí que se eliminen todos los pasajes que no estén presentes en ambas versiones o que no pertenezcan al cuerpo de la obra: información bibliográfica (que se almacena en un archivo aparte como metadatos), dedicatorias, epígrafes, prefacios, apéndices, notas o bibliografías.

A continuación, los textos son revisados cuidadosamente a fin de detectar eventuales errores provocados por el proceso de digitalización de las obras más antiguas, pues las más recientes ya han sido creadas digitalmente. Se marca la estructura interna de la obra, esto es, la eventual división en partes o capítulos. Por último, se recogen y almacenan en un archivo aparte los metadatos descriptivos de cada obra (*cfr*. Doval 2018: 183).

El alineado es, obviamente, una fase crucial en la construcción de un corpus paralelo, y es definido por Tiedemann (2011: 123) como «a process of making symmetric correspondences explicit in order to enable further processing of parallel resources». Este conjunto de correspondencias entre el texto fuente y el meta es lo que forma el llamado bitexto. Las unidades de alineado del bitexto dependen del nivel de detalle de la segmentación que se considere: parágrafos, oraciones o palabras. Actualmente en los recursos paralelos el alineado a nivel de oración constituye un estándar y es el más común en los corpus paralelos (Tiedemann 2011: 37). Por ello, en PaGeS nos hemos centrado en la oración como la unidad básica de alineado.

En este proceso se combinan dos tareas, una previa de segmentación de los textos en oraciones y la posterior vinculación de esos segmentos con sus correspondencias en la otra lengua para formar bisegmentos. La segmentación se lleva a cabo para cada lengua independientemente y el algoritmo realiza en principio un corte por cada puntuación fuerte, esto es, punto, signo de interrogación o de admiración final[4] (Zanettin 2012: 158). Ahora bien, como la puntuación en alemán y español no siempre es coincidente, esta segmentación inicial es posteriormente rectificada, caso necesario, en el proceso de alineado (*vid. infra*). Además, el signo de puntuación punto (.) es ambiguo, ya que no siempre indica el final de una oración, sino que se utiliza como marcador de abreviaturas (en alemán y español) y en alemán también como marcador de número ordinal, tal como ilustra el siguiente ejemplo:

(1) Dr. Kaltensee feiert heute in U.S.A. seinen 30. Geburtstag. [El Dr. Kaltensee celebra hoy en USA su 30º cumpleaños]

[4] Más precisamente (Scott 2010, citado en Zanettin 2012: 158) puntualiza: «A sentence ends if a full-stop, question-mark or exclamation-mark (.?!) is immediately followed by one or more word separators and if the next non-punctuation symbol is a capital letter A..Z or an accented capital letter, a number or a currency symbol».

Se hace, por tanto, necesario desambiguar el punto de final de oración del que forma parte de la palabra, como en las abreviaturas o en los números ordinales. Para ello, se integra una lista finita de abreviaturas comunes para alemán y español en el segmentador. Evidentemente, esta lista no es exhaustiva, ya que continuamente se crean abreviaturas nuevas u ocasionales.

Tiedemann (2011: 9) subraya la importancia de la segmentación para la precisión del alineado posterior: «The importance of segmentation is often ignored in the literature on text alignment. However, it plays a crucial role in the success of the algorithm». Efectivamente, buen número de los fallos en el alineado se deben a fallos en la segmentación previa. Esto ocurre especialmente en los casos de puntuación no coincidente, tal como ilustra el ejemplo siguiente, en el que en español hay dos puntos y no se han segmentado, mientras que en alemán hay un punto y se ha introducido, por tanto, una segmentación (Pérez Reverte 2012, cap. 3).

	Es kostete nichts, freundlich zu sein.
No costaba nada ser amable: invertir de cara al futuro.	Es war eine Investition in die Zukunft.

Tabla 3. Ejemplo de puntuación divergente

Tras realizar varios tests con diferentes herramientas de alineado[5], hemos optado por LF-Aligner, ya que alcanzó en ellos una mayor precisión. Está basado en el algoritmo de Hun-Align[6] (Varga 2012: 92-119), un enfoque híbrido que combina el algoritmo basado en la semejanza de longitud de las oraciones con el de las correspondencias léxicas (Tóth *et al.* 2008). Como las correspondencias léxicas se derivan de forma automática, el LF-Aligner no requiere de un léxico externo.

El alineado de estos segmentos resultantes se ejecuta en cuatro fases: 1) el archivo input es un texto «tokenizado» y segmentado en oraciones en las dos lenguas; 2) este texto se alinea usando una versión modificada del modelo de Brown *et al.* (1993), basado en la longitud de las respectivas oraciones; 3) se construye un diccionario automático bilingüe basado en este primer alineado y 4) finalmente, se realinea el texto en un segundo paso, tomando en consideración la semejanza léxica, proporcionada por el diccionario automático (Toth *et al.* 2008: 470).

[5] Se han realizado tests con los siguientes alineadores; ABBY Aligner, bitext2tmx y Vanilla aligner (Danielsson & Ridings 1997).

[6] Hun-Align es una elección común entre los creadores de corpus paralelos multilingües. Por ejemplo se usó para el alineado de Intercorp (Čermák 2019), en la plataforma Multilingwis (Clematide *et al.* 2016), o en el alineado de JRC-Acquis (Steinberg *et al.* 2014: 1) entre otros.

La precisión del alineado depende, en primer lugar, del tipo de textos; concretamente, los textos ficcionales, que forman el núcleo del corpus PaGeS, presentan mayores problemas que otro tipo de textos, como los administrativos o técnicos, tal como señala Zanettin (2012: 155):

> Some text types are better suited to automatic alignment than others. Typically, parallel collections of technical, legal or otherwise official documents such as transcriptions of parliamentary proceedings, especially in related languages, are, in this respect, much better material than journalistic and fictional texts. While the former usually present the same formal structure and contain the same number of paragraphs and sentences as well as fixed and standard punctuation, the latter often score low on these features.

Además, dentro de los textos literarios, el grado de correspondencia varía dependiendo del autor, del traductor, de los propios textos y de la dirección de la traducción.

Obviamente, una correspondencia 1:1 entre la lengua fuente y la meta no es siempre posible, puesto que en el proceso traductológico las oraciones pueden ser divididas (1:2), fusionadas (2:1) o reordenadas; además, el traductor puede optar por omitir o insertar oraciones o pasajes de texto (Tiedemann 2011: 9; Varga 2012: 94). Las tablas 4 y 5 muestran algunos de estos casos. Además, tal como se ha señalado anteriormente, PaGeS incluye también bitextos en los que ambas versiones alemana y española son traducciones de una tercera lengua y estos textos son particularmente problemáticos, ya que han sufrido dos procesos de traducción independientes (Doval 2018: 186).

	Im hinteren Teil sorgt eine Reihe von Leuchten für gleichmäßiges, gedämpftes Licht.
Al fondo, frente a ocho filas de asientos ocupados por el público, una instalación de luz artificial amortiguada, uniforme, ilumina la mesa de juego situada en una tarima y un gran tablero mural de madera que hay en la pared, junto a la mesa del árbitro, donde un ayudante de éste reproduce el desarrollo de la partida.	Auf einem Podest steht dort der Spieltisch vor acht mit Zuschauern besetzten Stuhlreihen, und an der Wand hängt ein großes Schachbrett, auf dem der Schiedsrichterassistent die Spielzüge nachstellt.

Tabla 4. Ejemplo de mal alineado debido a una correspondencia 1:2

	Sie lachte laut auf.
Ella soltó una carcajada viva y fuerte.	Ein fröhliches und herzhaftes Lachen.
Una risa sana. —Exacto —asentía, siguiéndole la corriente con buen humor—.	„Genau«, bestätigte sie. „Wie haben Sie das erraten?«

Tabla 5. Ejemplo de mal alineado debido a reordenación

Después del alineado automático, como se ha dicho, se valida manualmente el resultado. Solo de esta manera se puede conseguir una tasa de error inferior al 0,5 %. Para ello se procede en tres fases. Primero se seleccionan los segmentos de más de 350 caracteres, ya que han de ser divididos para poder ser procesados. Para ello se insertan manualmente marcas (breaks)
 en lugares adecuados del texto de ambos segmentos, que luego son divididos automáticamente. En un segundo paso se localizan los alineados vacíos, es decir, los segmentos no emparejados. En este caso, puede tratarse de un alineado erróneo o de eliminaciones o inserciones en el texto traducido. Si el segmento está desalineado, se hacen las correcciones necesarias. Si el segmento no ha sido traducido, se inserta en la celda vacía la marca [n_t_s] (=non translated segment). Si se ha añadido el segmento en el texto traducido, se inserta la marca [a_s_t] (=texto añadido en la traducción). Finalmente, para minimizar el trabajo manual, nos centramos en los bisegmentos en los que, debido a desfases de longitud entre el segmento fuente y el meta, es más probable que haya errores. Para identificarlos calculamos el cociente de la suma de caracteres del bisegmento y la diferencia de caracteres entre el segmento fuente y meta. Luego aplicamos esta ratio para ordenar los bisegmentos. Los errores tienden a ocurrir en los bisegmentos donde el rango de valores de la ratio está entre -5 - 5. De esta manera la comprobación manual de los resultados del alineado se realiza de forma más eficiente y requiere menos tiempo. Este procedimiento es un compromiso entre lo que sería deseable y lo que es factible, asegurando asimismo un alto nivel de precisión.

4. Búsqueda y visualización de los resultados

Una vez que los archivos alineados han sido revisados, se indexan y se gestionan a través del motor de búsqueda general Apache-Solr (Versión 7.5.0), una potente y muy rápida plataforma de búsqueda de código abierto escrita en Java, que cubre una amplia gama de funcionalidades.

Como se mencionó antes, PaGeS pretende ser una herramienta realmente polivalente, útil para grupos de usuarios muy diversos, desde investigadores en lingüística y traducción hasta lexicógrafos, expertos en PLN, pasando por usuarios no especialistas, sean ocasionales o habituales, así como estudiantes de alemán o español. Para ello es esencial proporcionarles una interfaz adecuada para la visualización y recuperación de los datos, esto es, los textos del corpus, los metadatos y las eventuales anotaciones lingüísticas. Para lograr este objetivo, la interfaz debe ofrecer una serie de funcionalidades básicas (Doval *et al.* 2019: 115). Por un lado, la búsqueda ha de ser:

(a) Rápida, ya que se prevé un aumento significativo del tamaño del corpus PaGeS. El motor de búsqueda debe permitir realizar búsquedas de forma rápida y eficaz a través de grandes cantidades de datos lingüísticos.
(b) Amigable. El lenguaje de la consulta debe ser lo más sencillo posible. Un lenguaje de consulta avanzado y más complejo solo se muestra si es necesario. Además, deben aprovecharse los hábitos de búsqueda de los usuarios de Internet y, por lo tanto, el lenguaje de consulta de Google debe servir de modelo.
(c) A varios niveles: El sistema de búsqueda debe permitir consultas a través de múltiples capas de anotación lingüística, como la lematización, el etiquetado de clases de palabras o la anotación sintáctica.

Por otro lado, los resultados de las consultas han de presentarse en un formato fácil de leer. Los segmentos fuente y meta deben mostrarse uno al lado del otro, y tanto la palabra o frase de búsqueda como su equivalente potencial deben ser resaltados.

Por ello y para satisfacer las necesidades de los usuarios mencionados anteriormente, hemos diseñado una búsqueda en tres niveles. La primera, cuya interfaz se muestra en la figura 3, es la búsqueda simple o estándar. En este caso, el usuario solo tiene que introducir en el campo de búsqueda el término de la búsqueda (una palabra o una frase) en alemán o español. En este tipo de consultas, la lematización se aplica por defecto. Con las consultas multipalabra, se encuentran todas las palabras de búsqueda dentro de una distancia específica. Al igual que en la búsqueda de Google, si el término se introduce entrecomillado (por ejemplo «pasar de largo») la búsqueda solo devuelve resultados que coincidan exactamente con la forma de la palabra o frase introducida (*vid*. figura 4). Por defecto, las búsquedas solo incluyen el corpus nuclear, aunque el usuario puede seleccionar la casilla correspondiente a los textos complementarios (Europarl o TED).

La forma más popular de mostrar los resultados de una búsqueda en un corpus es en forma de concordancia y el formato de concordancia más común es la concordancia KWIC (Key Word in Context) (Wynne 2008: 706), es decir, la palabra objeto de la consulta está en una posición central con todas las líneas alineadas verticalmente alrededor de ella. Esta presentación de los resultados, debidamente ordenada, es muy útil en corpus monolingües para visualizar patrones de uso. Sin embargo, en los corpus bilingües, el formato KWIC no puede considerarse amigable, ya que una de las principales aplicaciones de estos corpus es encontrar rápidamente posibles equivalentes

de un término de búsqueda. La figura 2 ilustra este tipo de presentación en un corpus alineado por AntPConc.

FIGURA 2. Visualización kwic de AntPConc

Por esta razón, nos decidimos por la visualización de los resultados de la consulta en una tabla de dos columnas, donde una columna corresponde a los textos fuente y la otra a los textos meta. El término de búsqueda se muestra en una celda con algún contexto y se resalta en negrita. Dependiendo de si el término de la búsqueda se encuentra en el texto original o en la traducción, se muestra en una u otra columna. Los resultados en los textos originales se muestran primero en la columna izquierda, los de las traducciones en la columna derecha. El segmento correspondiente se visualiza en la misma línea, pero en una celda de la otra columna. En cada consulta, se informa sobre el número de ocurrencias y el número de páginas, así como el número de la página actual. Las figuras 3 y 4 proporcionan ejemplos del menú de búsqueda estándar y algunas de sus características:

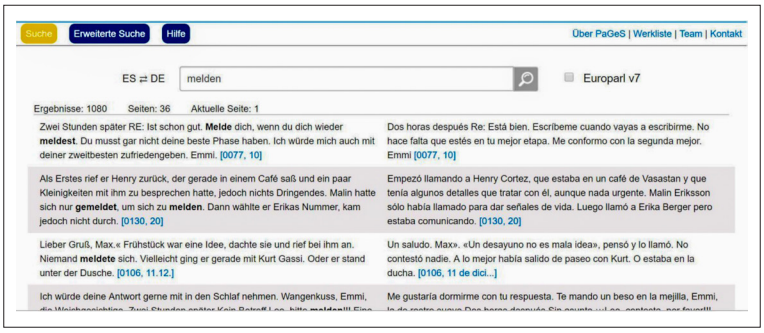


FIGURA 3. Búsqueda simple en el corpus PaGeS

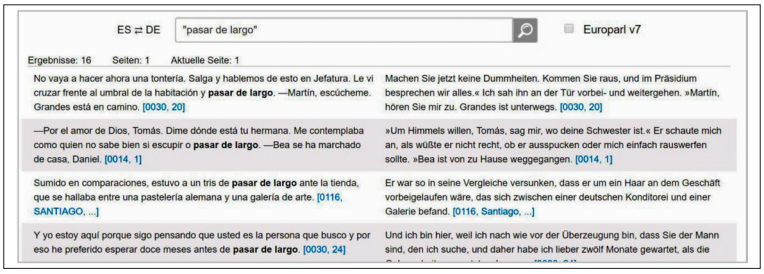


FIGURA 4. Búsqueda multipalabra exacta en el corpus PaGeS

Al final de la tabla hay un link que permite navegar a través de las páginas y descargar los resultados de la búsqueda en formato CSV a los usuarios registrados En cada pasaje se ofrece, entre corchetes, información relativa al ID de la obra, así como la parte y el capítulo en el que se encuentra. Haciendo clic sobre el ID de la obra, el usuario puede ver un contexto lingüístico mayor, seleccionando el número de segmentos antes y después de la ocurrencia. Además, en esta pantalla se muestra la información bibliográfica completa de la obra, tal como muestra la figura 5.

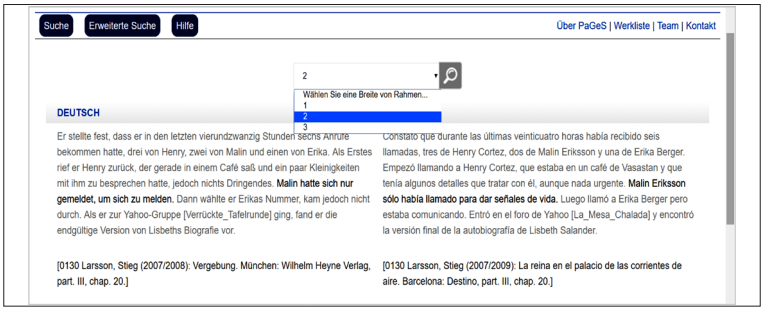


FIGURA 5. Ampliación de contexto e información bibliográfica

El segundo nivel de búsqueda es el avanzado. En este nivel, el usuario puede restringir el alcance de su búsqueda aplicando filtros operativos mediante menús desplegables. Puede limitar la búsqueda por autores, obras, años de publicación o género. También puede seleccionar si la búsqueda ha de operar solo en originales o en traducciones. Además, puede buscar un término solo cuando se traduce de una manera concreta. La figura 6 muestra un ejemplo de búsqueda avanzada.

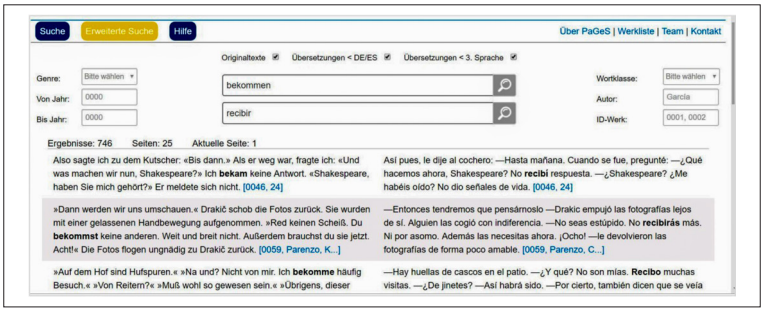


FIGURA 6. Búsqueda avanzada en PaGeS

El último nivel de búsqueda es el más complejo, y actualmente está parcialmente disponible a través de la interfaz de búsqueda estándar. Este nivel proporciona al usuario el comando completo de la sintaxis de consulta utilizada en la herramienta de indexación subyacente Solr. Esta soporta búsquedas usando expresiones regulares. El término de búsqueda debe ir precedido de [SS] (Solr Search). La expresión de búsqueda ha de ser construida palabra a palabra, y para cada palabra, es posible especificar varios parámetros. Este nivel permitirá, además, la búsqueda combinada de palabras o frases y etiquetas de clases de palabras u otras anotaciones, cuando estén implementadas. Esta búsqueda formal permite la ejecución de consultas muy precisas, pero no es particularmente fácil de usar. Está dirigida a usuarios más exigentes, como los investigadores en lingüística contrastiva o traducción, que suelen necesitar un subconjunto muy específico de resultados, solo posible con consultas complejas que incluyan un gran número de parámetros.

5. BALANCE Y PERSPECTIVAS

En los apartados anteriores hemos presentado las características distintivas del corpus PaGeS y los procesos implicados en su compilación e indexación. Los tres niveles de búsqueda están adaptados a las necesidades de los distintos grupos de usuarios, lo que hace del corpus PaGeS un recurso multifuncional con un enorme potencial. Sus aplicaciones concretas en la lingüística contrastiva y el aprendizaje de idiomas se están explotando actualmente en nuestro grupo de investigación (Doval 2018, Lübke & Liste Lamas 2018).

En todo caso tenemos previsto introducir nuevas funcionalidades. Las más inmediatas y que ya están en parte realizadas son el alineado de palabras y el etiquetado de clases de palabras. El alineado de palabras es muy útil, especialmente para los estudiantes, ya que facilita la identificación a primera

vista de la palabra equivalente del término de búsqueda en la otra lengua (Volk *et al.* 2014). Actualmente se están realizando pruebas con los siguientes alineadores: Giza++, Berkeley Aligner y eflomal.

En cuanto al etiquetado con clases de palabras ya ha sido realizado para el corpus nuclear, utilizando para el español el etiquetador Freeling y para el alemán el TreeTagger (*vid.* Doval 2017a). La inclusión de esta información en el indexado permitirá realizar consultas complejas, en las que se combine la búsqueda de la cadena y la categoría.

Por otro lado, ya ha sido iniciada la construcción de otros dos corpus paralelos: español/holandés, el corpus PaDeS, y español/chino, el corpus PaZheS, con los mismos criterios de diseño y características de consulta. No están todavía en línea, pues no disponen en el momento actual de la suficiente cantidad de material paralelizado mínimo para que pueda resultar útil. Está prevista su puesta en línea en cuanto se alcancen los cinco millones de palabras en el corpus nuclear.

Por último, somos muy conscientes de algunas deficiencias actuales de nuestro motor de búsqueda, como la limitación en la ordenación de los resultados o la selección de colocaciones. De todos modos y, a pesar de la existencia de otros corpus paralelos, PaGeS presenta una serie de características distintivas, entre las que cabe destacar: el tipo de textos utilizado de gran variedad léxica y gramatical, la alta calidad de originales y traducciones, el equilibrio en la bidireccionalidad, la revisión manual de los procesos automáticos y, finalmente, su disponibilidad en línea y su facilidad de uso. Todas estas características lo convierten en un recurso multifuncional adecuado para múltiples aplicaciones, desde la investigación contrastiva, pasando por los estudios de traducción hasta el aprendizaje y enseñanza de lenguas extranjeras.

Recursos electrónicos

ABBY Aligner: https://www.abbyy.com/en-eu/aligner/

AntPConc: https://www.laurenceanthony.net/software/antpconc/

Apache-Solr (Versión 7.5.0) http://lucene.apache.org/solr/

Berkeley Aligner: https://github.com/mhajiloo/berkeleyaligner

Bitext2tmx http://bitext2tmx.sourceforge.net/

Eflomal: Efficient Low-Memory Aligner https://github.com/robertostling/eflomal

Freeling: http://nlp.lsi.upc.edu/freeling/node/1

Giza++: https://github.com/moses-smt/giza-pp

LF-Aligner: http://sourceforge.net/projects/aligner/

Linguee: Linguee GmbH (2010): Linguee.com: The web as a dictionary. http://www.linguee.com

Multilingwis: https://pub.cl.uzh.ch/projects/sparcling/multilingwis2.demo/

Opus: http://opus.nlpl.eu/

TED: www.ted.com

TreeTagger: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

WIT3: Web Inventory of Transcribed and Translated Talks https://wit3.fbk.eu/

Referencias bibliográficas

Aijmer, Karen, Bengt Altenberg & Mats Johansson (eds.) (1996): *Languages in contrast: papers from a symposium on text-based cross-linguistic studies in Lund, 4-5 March 1994*. Lund: Lund University Press, pp. 73-85.

Borin, Lars (2002): «...and never the twain shall meet?» in Lars Borin (ed.): *Parallel corpora, parallel worlds: selected papers from a symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22-23 April, 1999*. Amsterdam: Rodopi, pp. 1-43. https://doi.org/10.1163/9789004334298_002

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra & Robert L. Mercer (1993): «The mathematics of statistical machine translation: parameter estimation», *Computational Linguistics* 19/2, pp. 263-311.

Čermák, Petr (2019): «InterCorp: Parallel corpus of 40 languages», in Irene Doval & M. Teresa Sánchez (eds.): *Parallel corpora for contrastive and translation studies: new resources and applications*. Amsterdam: John Benjamins, pp. 93-101. https://doi.org/10.1075/scl.90.06cer

Cettolo, Mauro, Christian Girardi & Marcello Federico (2012): «WIT3: Web Inventory of Transcribed and Translated Talks», in *Proceedings of EAMT*, Trento, Italy, pp. 261-268.

Clematide, Simon, Johannes Graën & Martin Volk (2016): «Multilingwis: a multilingual search tool for multi-word units in multiparallel corpora», in Gloria Corpas Pastor (ed): *Computerised and corpusbased approaches to phraseology: monolingual and multilingual perspectives – Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües*. Geneva: Tradulex, pp. 447-455.

Doval, Irene & M. Teresa Sánchez Nieto (2019): «Parallel corpora in focus: an account of current achievements and challenges», in Irene Doval & M. Teresa Sánchez Nieto (eds.): *Parallel corpora for contrastive and translation studies: new resources and applications*. Amsterdam: John Benjamins, pp. 1-19. https://doi.org/10.1075/scl.90.01dov

Doval, Irene, Santiago Fernández Lanza, Tomás Jiménez Juliá, Elsa Liste Lamas & Barbara Lübke (2019): «Corpus PaGeS: a multifunctional resource for language learning, translation and cross-linguistic research», in Irene Doval & M. Teresa Sánchez Nieto (eds.): *Parallel corpora for contrastive and translation studies: new resources and applications*. Amsterdam: John Benjamins, pp. 103-121. https://doi.org/10.1075/scl.90.07dov

Doval, Irene (2017a): «POS-tagging a bilingual parallel corpus: methods and challenges», *Research in Corpus Linguistics* 5, pp. 35-46. https://doi.org/10.32714/ricl.05.03

Doval, Irene (2017b): «La construcción de un corpus paralelo bilingüe multifuncional», *Moenia: Revista lucense de lingüistica & literatura* 27, pp. 125-141.

Doval, Irene (2018): «Das PaGeS-Korpus, ein Parallelkorpus der deutschen und spanischen Gegenwartssprache», *Revista de Filología Alemana* 26, pp. 181-197. https://doi.org/10.5209/RFAL.60148

Hasselgård, Hilde (2015): «Parallel corpora and contrastive studies», in *Proceedings of the international symposium on Using Corpora in Contrastive and Translation Studies, (UCCTS), 2010 Conference,* Edge Hill University, 27-29 July 2010. http://www.lancaster.ac.uk/fass/projects/corpus/UCCTS2010/Proceedings/papers/Hasselgard. pdf [consultado: 26 de junio de 2018].

Johansson, Stig & Knut Hofland (1994): «Towards an English-Norwegian parallel corpus», in Udo Fries, Gunnel Tottie & Peter Schneider (eds): *Creating and using English language corpora*. Amsterdam: Rodopi, pp. 25-37.

Koehn, Philipp (2005): «Europarl: a parallel corpus for statistical machine translation», in *Proceedings of Machine Translation Summit X*, Phuket, 13-15 September 2005. Vol. 5, pp. 79-86.

Lübke, Barbara & Elsa Liste Lamas (2018): *Raumrelationen im Deutschen: Kontrast, Erwerb und Übersetzung*. Tübingen: Stauffenburg.

McEnery, Tony & Andrew Hardie (2012): *Corpus linguistics: method, theory and practice*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

Steinberger, Ralf, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco Benítez, Patrick Schlüter, Marek Przybyszewski & Signe Gilbro (2014): «An overview of the European Union's highly multilingual parallel corpora», *Language Resources and Evaluation* 48/4, pp. 679-707. https://doi.org/10.1007/s10579-014-9277-0

Tiedemann, Jörg (2011): *Bitext alignment*. [*s. l.*]: Morgan & Claypool Publishers. https://doi.org/10.2200/S00367ED1V01Y201106HLT014

Tiedemann, Jörg (2012): «Parallel data, tools and interfaces in OPUS», in *Proceedings of the 8th International Conference on Language Resources*

and Evaluation (LREC '2012), pp. 2214-2218. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf [consultado: 20 de mayo de 2018].

Tóth, Krisztina, Richárd Farkas & András Kocsor (2008): «Sentence alignment of Hungarian-English parallel corpora using a hybrid algorithm», *Acta Cybern* 18, pp. 463-478.

Varga, Dániel (2012): *Natural language processing of large parallel corpora*. PhD thesis. Eötvös Loránd University, Budapest.

Volk, Martin, Johannes Graen & Elena Callegaro (2014): «Innovations in parallel corpus search tools», in *Proceedings of LREC, Reykjavik*. http://www.zora.uzh.ch/id/eprint/97282 /1/Volk_Graen_Callegaro_LREC_2014_v06.pdf

Wynne, Martin (2008): «Searching and concordancing», in Anke Lüdeling & Merja Kytö (eds): *Corpus linguistics: an international handbook*. Berlin: de Gruyter, pp. 706-737.

Xiao, Richard (2010): *Using corpora in contrastive and translation studies*. Cambridge Scholars Publishing.

Zanettin, Federico (2012): *Translation-driven corpora: corpus resources for descriptive and applied translation studies*. London: Routledge.


USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA