ISSN: 2174-7245
Año 2025, Volumen 15, Número 1.doi: 10.7203/Normas.v15i1.31001
Publicado: 2025. Enviado: 2024-05-08. Aceptado: 2025-06-25.

Corpus oral para el estudio de las relaciones entre lengua y género en español. El corpus CoLaGe

Oral corpus for the study of language and gender relations in Spanish. The CoLaGe corpus

Andrea Carcelén Guerrero
U. of Helsinki

Pekka Posio
U. of Helsinki

Sven Kachel
U. of Kassel | U. of Helsinki

Gloria Uclés
U. d'Alacant | U. of Helsinki

Abstract

The aim of this paper is to present and describe the particularities of creation and design of the oral corpus of Spanish CoLaGe (Corpus for the study of Language and Gender in Mexico and Spain), a bidialectal corpus obtained in Valencia (Spain) and Guadalajara (Mexico). For its elaboration, linguistic data (sociolinguistic interview, role play and image description task with phonetic purposes) and socio-psychological data have been collected for the study of the relationships between the gender of the speakers and their use of the language, in order to deepen the understanding of the relationship between language, society and gender. The corpus includes 127 informants, one million words and more than 100 hours of recordings.

Keywords: oral corpora, linguistic variation, sociolinguistics, gender, corpus linguistics.

Resumen

Este trabajo tiene como objetivo presentar y describir las particularidades de creación y diseño del corpus oral del español CoLaGe (Corpus for the study of Language and Gender in Mexico and Spain), un corpus bidialectal obtenido en Valencia (España) y Guadalajara (México). Para su elaboración se han recogido datos de carácter lingüístico (entrevista sociolingüística, juego de rol y tarea de descripción de imágenes con propósitos fonéticos) y datos sociopsicológicos para el estudio de las relaciones entre el género de los hablantes y su uso de la lengua, de manera que se profundice en la comprensión de la relación entre lengua, sociedad y género. El corpus cuenta con 127 informantes, un millón de palabras y más de 100 horas de grabaciones.

Palabras clave: corpus orales, variación lingüística, sociolingüística, género, lingüística de corpus.

Citar como: Carcelén Guerrero, Andrea; Posio, Pekka; Kachel, Sven y Uclés Ramada, Gloria (2025). Corpus oral para el estudio de las relaciones entre lengua y género en español. El corpus CoLaGe. Normas, 15(1), 1-19, doi: 10.7203/Normas.v15i1.31001.

1. Introducción[1]

El panorama de corpus orales del español se encuentra en un momento álgido; así lo demuestran los trabajos recientes de Carcelén (2024), Briz y Samper (2022) y Llisterri (2021) que ofrecen una perspectiva general de la situación actual de los corpus orales para esta lengua; una panorámica rica en muestras de habla orales disponibles para el análisis de la variación lingüística.

Si analizamos las características en común de los trabajos de corpus recogidos por los autores mencionados, podemos observar que la mayoría obtienen sus muestras de habla siguiendo criterios de estratificación sociolingüística. Esto es, los participantes han sido seleccionados atendiendo a criterios como el grupo etario, el sexo (hombre-mujer) y el nivel sociocultural (alto, medio, bajo). Véanse, por ejemplo, los criterios de estratificación seguidos por el corpus de la Real Academia Española y la Asociación de Academias de la Lengua Española, el *Corpus del español del siglo XXI* (Corpes XXI), el corpus del *Proyecto para el Estudio Sociolingüístico del Español de España y América* (PRESEEA), el corpus Ameresco (América y España Español Coloquial) o el *Corpus del Español Contemporáneo de México* (CECM).

La particularidad del corpus que presentamos aquí radica en que, además de establecer el criterio basado en sexo, recoge otros datos de los informantes que van más allá de la clasificación binaria hombre-mujer, apostando por adoptar un enfoque escalar en cuanto al autoconcepto del rol de género[2] que amplía la perspectiva de estudio al contemplar el grado de masculinidad o feminidad al que se adhieren los informantes, medidos con diferentes escalas psicológicas basadas en la autoevaluación de los rasgos de personalidad y sus comportamientos.

Este trabajo tiene como objetivo presentar el proceso de diseño y construcción del corpus CoLaGe a través de las siguientes secciones: en el apartado 2 se da cuenta de las particularidades del proyecto de investigación en el que se enmarca el corpus, el proyecto *Gender, Society and Language Use: evidence from Mexico and Spain*; en el apartado 3 se detalla la metodología de recogida del corpus (subapartado 3.1.) y el protocolo de tratamiento de los datos recopilados (subapartado 3.2.). Por último, a modo de conclusión, en el apartado 4, se recogen las consideraciones finales.

2. El proyecto *Gender, Society and Language Use: evidence from Mexico and Spain*

El proyecto *Gender, Society and Language Use: evidence from Mexico and Spain* (Pekka Posio, dir.) surge en la Universidad de Helsinki, financiado por la Fundación Kone, y está formado por especialistas en el ámbito de la lingüística y la psicología procedentes de diferentes universidades internacionales.

Este proyecto tiene como objetivo estudiar las interrelaciones entre el género de los hablantes y su uso de la lengua, de manera que se profundice en la comprensión de la relación entre lengua, sociedad y género. En particular, pretende investigar de forma comparada estas relaciones

[1]Esta investigación ha sido financiada por una beca de la Fundación Kone (202007066) en la Universidad de Helsinki.

[2]Entendemos *género* como un concepto que incluye componentes psicológicos, sociológicos y sociolingüísticos, más allá de la distinción biológica de sexo. Véase Prince (1985).

en dos zonas lingüísticas, en Valencia-España y en Guadalajara-México. A saber, pretende analizar qué papel juega el género y el sexo de los hablantes a la hora de explicar la variación lingüística.

Es por ello por lo que se plantea averiguar si las diferencias en el uso de la lengua se explican mejor a través del establecimiento de un enfoque escalar en cuanto al autoconcepto del rol de género, a diferencia del enfoque tradicional que considera el binomio *hombre/mujer* como categorías binarias y, por tanto, mutuamente excluyentes (Bem, 1974, Thompson & Pleck, 1986, Kachel *et al.*, 2016).

La cuestión fundamental que se aborda, por tanto, es hasta qué punto la sociedad afecta a los roles de género de los hablantes y a las particularidades de género en su habla, así como de qué manera se refleja este hecho en el español hablado en dichas ciudades. Con este fin, se han recogido dos corpus orales de base lingüística y sociopsicológica que comparan las dos zonas dialectales mencionadas (CoLaGe-V y CoLaGe-G). Además, se ha recopilado un subcorpus exploratorio en Guadalajara (CoLaGe-G-Diversity) obtenido de informantes pertenecientes a diferentes minorías de género y/u orientación sexual.

3. El corpus CoLaGe (*Corpus for the study of Language and Gender in Mexico and Spain*)

El corpus CoLaGe (Uclés *et al.*) nace con la finalidad de ser la herramienta fundamental de análisis para las hipótesis y objetivos planteados en el seno del proyecto. Pero, además, no solo da respuesta a las necesidades particulares de investigación de este, sino que pretende aumentar la nómina de corpus orales del español disponibles en abierto para toda la comunidad científica, de manera que pueda servir de base de datos para otros intereses de investigación.

A continuación, presentamos las particularidades metodológicas de su diseño y construcción atendiendo a las diversas fases de trabajo. Por un lado, trataremos sobre la metodología de recogida de los datos y, por otro, sobre el protocolo de tratamiento de los materiales obtenidos.

3.1. Metodología de recogida de los datos

Los intereses de investigación de este proyecto están centrados en el estudio de dos variedades del español, en concreto, del español de Valencia (España) y el de Guadalajara (México). Ambas localizaciones tienen en común que son grandes urbes con una gran presencia de actividad industrial y vida universitaria. No obstante, difieren en sus características sociológicas, esto es, según hipótesis iniciales, la sociedad española se presenta como menos conservadora que la mexicana; así lo afirman informes como Hausmann *et al.* (2014), Social Watch (2012), INEGI (2016) y la Comisión Europea (2017), que señalan diferencias notorias en cuanto a las normas de género, siendo la sociedad mexicana más conservadora que la española.

Para la obtención de los datos se ha contado con dos grupos colaboradores[3], en la Universidad de Guadalajara, para el caso de CoLaGe-G y CoLaGe-G-Diversity, y en la Universitat de

[3]Con la colaboración de Patricia Córdova (Universidad de Guadalajara) y Marta Albelda (Universitat de València) y de personal procedente de ambas universidades, encargado de grabar, transcribir y codificar, bajo la supervisión del equipo central de la Universidad de Helsinki.

València, para CoLaGe-V. Las grabaciones de los corpus se recogieron entre los años 2022 (CoLaGe-V) y 2023 (CoLaGe-G y CoLaGe-G-Diversity).

Selección de los informantes y tamaño de la muestra

Para la selección de los hablantes se han fijado unas cuotas estratificadas que contemplan diversas clasificaciones, como puede observarse en la Tabla 1.

Por un lado, los participantes pertenecen a dos grupos etarios distintos, con edades comprendidas entre los 30-40[4] para el grupo 1 y entre los 60-70 para el grupo 2. El motivo de esta división tiene que ver con la hipótesis inicial en la que se espera que el grupo comprendido entre los 30-40 presente posturas menos conservadoras y más abiertas con respecto a la segunda generación que, esperablemente, sería más tradicional. Así, los informantes del grupo 2 han nacido en periodos marcados por regímenes políticos opresivos, y los informantes del grupo 1 se han desarrollado en una sociedad más abierta y democrática.

Corpus	**Edad**	**Sexo**			**Total**
		Mujer	**Hombre**	**No binario**	
CoLaGe-V	30-40	14	13	0	27
	60-70	12	12	0	24
	Total	**26**	**25**	**0**	**51**
CoLaGe-G	30-40	15	15	0	30
	60-70	15	15	0	30
	Total	**30**	**30**	**0**	**60**
CoLaGe-G-Diversity	30-40	**5**	**2**	**0**	8
	60-70	**3**	**5**	**1**	8
	Total	**8**	**7**	**1**	**16**
Total		**64**	**62**	**1**	**127**

Tabla 1. *Resumen de la muestra extraída*

Por otro lado, para la variable sexo se ha contemplado la división binaria entre mujer y hombre, añadiendo además, la opción de género no binario[5]. Esta estratificación se puede combinar con el grado de conformidad de género con el que se identifican los participantes, quienes a través del cuestionario sociopsicológico (ver apartado 3.1.3.2.) se han situado en

[4]En la muestra recogida en Valencia, el rango de edad finalmente se estableció entre los 29 y 41 para el grupo 1. Esta diferencia se debió a la dificultad para acceder a más informantes debido, en parte, a la situación pospandémica en la que se recogieron las grabaciones. En el momento de la recogida aún existían ciertas restricciones como la obligatoriedad de usar de mascarillas, límites en cuanto a la distancia social y las posibilidades de reunión en un mismo espacio de varias personas al mismo tiempo, condiciones que provocaron las reticencias de participación de posibles informantes.

[5]Además, se ha recogido información sobre la orientación sexual de los informantes, de modo que se pueda considerar como una variable de análisis complementaria.

un determinado punto de la escala en cuanto al autoconcepto de rol de género y autoimagen propuesta por Kachel *et al.* (2016). En esta gradación se contemplan las variables siguientes entre el extremo más femenino y el extremo opuesto más masculino:

1. *Me considero…*
2. *Idealmente, me gustaría ser…*
3. *Tradicionalmente, mis intereses se considerarían como…*
4. *Tradicionalmente, mis actitudes y creencias se considerarían…*
5. *Tradicionalmente, mi comportamiento se consideraría como…*
6. *Tradicionalmente, mi aspecto exterior se consideraría…*

Por último, se reclutaron informantes que representaran a hablantes de clase media con nivel educativo medio-alto; todos han cursado al menos estudios de bachillerato.

En total, el corpus cuenta con 127 informantes, 60 para Guadalajara, 51 para Valencia y 18 para el subcorpus exploratorio Diversity Guadalajara. Cada sesión de recogida de datos tiene una duración aproximada de 1 hora y 15 minutos, repartidos de la siguiente manera: 40 minutos para la entrevista, 10 minutos para el juego de rol, 10 minutos para la tarea fonética y 15 minutos para cumplimentar el cuestionario. El corpus cuenta con 82 horas de entrevistas grabadas, 16 horas de juegos de rol y 12 horas de datos grabados para el análisis fonético. En total, está compuesto un millón de palabras y más de 100 horas de grabaciones.

El consentimiento informado

Para la recolección de las grabaciones y el cuestionario sociopsicológico que conforman el corpus CoLaGe, los informantes tuvieron que rellenar y firmar un formulario de consentimiento informado en el que autorizan al proyecto al tratamiento y almacenamiento de los datos, cumpliendo la legislación vigente según el *Reglamento general de protección de datos de la Unión Europea 2016/679*, así como las particularidades de la legislación de cada uno de los territorios donde se recopiló el corpus.

Al iniciar la recogida, cada participante fue informado de los objetivos de investigación, de quiénes son el equipo y las personas responsables del tratamiento de los datos, así como de cuáles son sus derechos de actuación con respecto a los datos recopilados. Además, son conocedores de que los datos personales serán convenientemente pseudonimizados (ver subapartado 3.2.3) y de que serán publicados en repositorios institucionales con fines exclusivamente de investigación y de docencia.

Asimismo, el personal técnico encargado de recoger los materiales y de su procesamiento firmó un acuerdo de confidencialidad en el que se comprometieron a cumplir con las instrucciones de seguridad y privacidad con respecto al material recogido.

Con este protocolo se garantiza el cumplimiento de las diferentes legislaciones que aplican a la recogida de datos personales.

Materiales que componen el corpus

Dado que el objetivo del proyecto es estudiar las interrelaciones entre el sexo y el género de los hablantes y su uso de la lengua para profundizar en la comprensión de la relación entre lengua, sociedad y género, para la recolección de este corpus ha sido necesario establecer dos tipos de materiales: lingüísticos, datos orales emitidos por los informantes que han sido obtenidos por medio de diferentes tareas grabadas y transcritas (entrevista sociolingüística, juegos de rol y tarea de descripción de imágenes con fines fonéticos), y materiales sociopsicológicos, recopilados por medio de un cuestionario en línea.

Los materiales lingüísticos se obtuvieron consecutivamente mediante una grabación de audio no secreta, utilizando un sistema de grabación profesional a la vista de los participantes. Además, para la realización de la tarea de descripción de imágenes y el cuestionario sociopsicológico, los informantes utilizaron una *tablet* que les permitía leer y aceptar el formulario de consentimiento informado, ver las imágenes que debían describir para la tarea fonética y, finalmente, completar el cuestionario sociopsicológico. Este cuestionario se completó en línea a través de la plataforma SoSciSurvey (Leiner, 2019). Presentamos, a continuación, las particularidades de cada uno de ellos.

Materiales lingüísticos

La entrevista sociolingüística

La entrevista sociolingüística se plantea como un método de recogida de datos que garantiza la obtención de muestras de habla representativas socialmente y de buena calidad y que permite la observación del cambio lingüístico (Recalde y Vázquez, 2018:56).

La entrevista se ha organizado en dos partes: la primera mitad gira en torno a una temática general, mientras que en la segunda mitad la temática es específica, relacionada con los intereses de investigación del proyecto, esto es, activar el tema del género. En ambas partes se incluyen preguntas que dan pie a que los informantes produzcan secuencias descriptivas, narrativas y de opinión.

Si bien el esquema de preguntas ha sido elaborado específicamente por el proyecto, para la primera parte se han tomado como referencia modelos existentes como el del *Proyecto para el Estudio Sociolingüístico del Español de España y América* (PRESEEA). Los informantes fueron preguntados por cuestiones de temática cotidiana siguiendo la siguiente estructura:

Calentamiento

> *Saludos, pregunta sobre el tiempo, ¿te han entrevistado alguna vez?, ¿eres de Guadalajara/Valencia?, ¿desde hace cuánto vives aquí?*

Entrevista parte 1

> *-¿Podrías describir el barrio en el que vives? Por ejemplo, ¿tiene parques, tiendas, mercados, teatros, zonas para hacer ejercicio?*
>
> *-¿Cómo te mueves por la ciudad? ¿Usas vehículo particular, transporte público, bicicleta, caminando?*
>
> *-Durante la pandemia han venido menos turistas a Guadalajara/Valencia. ¿Qué te parece este cambio?*

-¿Crees que es algo positivo o negativo que Guadalajara/Valencia sea un destino turístico?

-Se dice que Guadalajara/Valencia es menos segura ahora que hace unos años, ¿tú qué piensas? ¿Es seguro caminar por la noche? ¿Usar el transporte público?

-¿Podrías contarme alguna situación donde hayas pasado miedo? Si no has pasado miedo nunca, ¿has experimentado algún momento muy incómodo? ¿Cómo fue?

-¿Podrías contarme cómo era un día normal para ti durante la cuarentena en 2020? ¿Estabas en casa? ¿Tenías que ir a trabajar? ¿Qué diferencias había con tu rutina anterior?

Para la segunda parte, los informantes respondieron a cuestiones relacionadas con el género, es decir, estas preguntas iban dirigidas a obtener qué ideas tenían sobre los estereotipos de género y roles sociales según el sexo y el género. La finalidad del cambio se relaciona con el objetivo de determinar de qué manera el cambio de tópico puede alterar el comportamiento lingüístico de los informantes. En esta parte, el guion de la entrevista se organizó de la siguiente forma:

Entrevista parte 2

-¿Cómo crees que tu apariencia, personalidad, aficiones, intereses se asemejan o se diferencian de lo que se suele esperar de un hombre/mujer?

-Hay gente que cree que en las relaciones se esperan cosas diferentes de hombres y mujeres. ¿Podrías describir lo que se espera de una mujer/hombre? ¿Podrías mencionar algún ejemplo que conozcas?

-¿Crees que hombres y mujeres reciben un trato igualitario en México/España? Por ejemplo, en el trabajo, en casa, en una relación amorosa.

-¿Quién crees que sufre más violencia, las mujeres o los hombres? ¿En qué situaciones crees que los hombres/las mujeres sufren más violencia?

- ¿Alguna vez has sentido que te han tratado diferente por el hecho de ser una mujer/un hombre? ¿Podrías contar qué pasó? ¿Por parte de tus padres? ¿En la escuela? ¿En el trabajo? ¿Con tus amigos?

-¿Podrías contarme alguna anécdota que te haya pasado durante una primera cita? ¿Pasó algo inesperado? ¿Te sorprendió algo? ¿Te sentiste decepcionado/a?

-En caso de que no haya tenido citas, ¿podrías describir cómo sería una primera cita ideal para ti?

- ¿Crees que hay diferencias en la forma en la que se comunican hombres y mujeres? Hay gente que dice que los hombres y las mujeres hablan de forma diferente. ¿A qué crees que se refieren con eso? ¿Podrías dar algún ejemplo?

-¿Qué opinas del lenguaje inclusivo? Por ejemplo, el uso de amigos *y* amigas *o el uso de* amigues*.*

Entre ambas partes, las personas encargadas de realizar la entrevista introducían preguntas de calentamiento y transición para ayudar a los informantes a familiarizarse con la situación.

Juego de rol

Para esta segunda tarea se eligió el juego de rol como método para la elicitación de datos. En este caso, se planteó a los informantes dos situaciones ficticias en las que se buscaba obtener situaciones comunicativas conflictivas. Como señalan Recalde y Vázquez (2018) y Albelda (2022), esta técnica permite obtener este tipo de secuencias dada la dificultad de acceder a situaciones de conflicto con otros métodos.

Para la ejecución de esta tarea, se diseñaron dos situaciones comunicativas diferentes para las que cada participante (entrevistador e informante) adoptaba un rol concreto y recibía unas pautas de actuación. El guion que debían seguir puede verse en las Tablas 2 y 3:

Situación 1	
Informante	**Entrevistador**
Imagina que tu amigo te ha pedido como un favor que lo lleves a otra ciudad, a 50 km de Guadalajara/Valencia, para hacer unos trámites administrativos. Tú y tu amigo pertenecéis al mismo grupo de amigos de la escuela. Lo conoces desde hace mucho tiempo, pero no es uno de los amigos más cercanos. Últimamente has estado considerando terminar la amistad, ya que tu relación con él parece un poco unilateral. Tu amigo solo se pone en contacto contigo cuando necesita un favor o desahogarse. Además, cuando hay dinero involucrado, siempre dice que no tiene, o se olvida de pagarte. Sin embargo, al mismo tiempo, parece tener dinero para gastar en vacaciones, salir a comer y otros caprichos caros. Cuando llegas a la ciudad, no hay plazas de aparcamiento disponibles. Tu amigo insiste en aparcar en un lugar reservado para personas con movilidad reducida, ya que tiene prisa. Afirma que no habrá problema, ya que nadie revisa las infracciones de aparcamiento. Tú no te sientes cómodo aparcando allí, pero te sientes presionado por tu amigo. Cuando vuelves, ves que hay una multa en el coche. Realmente no quieres pagarla, pero sugieres dividir la multa. • Has llevado a tu amigo a una ciudad 50 km ida y vuelta como un favor y ni siquiera le has pedido dinero para pagar la gasolina. • Tu amigo te ha presionado para aparcar allí cuando no querías hacerlo, porque tenía prisa. • Piensas que tu amigo está aprovechándose de ti, como ha pasado muchas veces antes. • Piensas que tu amigo no está asumiendo las consecuencias de sus acciones. • Tu amigo dice que no tiene dinero, pero recientemente se fue de vacaciones a París, donde se hospedó en hoteles caros.	Imagina que le pides a tu amigo que te lleve a otra ciudad. Tienes prisa y no hay aparcamiento disponible. Sugieres que tu amigo aparque en un sitio reservado para personas con movilidad reducida, ya que será solo por un rato y, de todos modos, nadie revisa esos sitios. Cuando vuelves, ves que hay una multa en el coche. Tu amigo sugiere que cada uno pague la mitad de la multa. No quieres pagar porque no estabas conduciendo y no es tu coche. Vas a evitar pagar a toda costa. • Te niegas a pagar porque no es tu coche y no aparcaste allí. • Nunca forzaste a tu amigo a aparcar en ese lugar, lo hizo de manera libre. Así que ahora tiene que aceptar las consecuencias de sus acciones. • Estás corto de dinero y no puedes permitirte un gasto inesperado. • En el pasado, has pagado muchas cosas por tu amigo y nunca le has pedido el dinero. • Piensas que tu amigo está siendo egoísta y poco razonable.

Tabla 2. *Guion para representar la situación 1 en el juego de rol*

Situación 2 **Informante**	**Entrevistador**
Imagina que vives en un apartamento y tienes un vecino ruidoso. Constantemente escuchas ruidos provenientes de su apartamento: música y televisión a volumen alto, gritos, movimiento de muebles, etc. A lo largo de los años, te has quejado varias veces a tu vecino. Sin embargo, aunque siempre dice que va a hacer menos ruido, sigue siendo igual de ruidoso. Aunque no te gusta la situación, la has aceptado, ya que te gusta tu apartamento y el vecindario, y eres propietario, por lo que no quieres mudarte. Actualmente, estás en medio de unas renovaciones en tu casa que tardarán dos meses en completarse. Tu vecino viene a tu apartamento a quejarse del ruido que están causando las obras. No te agrada que la persona que ha estado haciendo ruido constantemente ahora se moleste por las reformas. • Has soportado el ruido de tu vecino desde que se mudó al apartamento hace 6 años. • Les has pedido que sea más tranquilo después de las 10 p.m., pero nunca lo ha hecho. • Tu vecino es ruidoso no solo durante el día, sino también bien entrada la noche. • El ruido de las obras siempre se hace durante las horas permitidas, nunca por la noche o durante el fin de semana. • Tienes una licencia emitida por el ayuntamiento para hacer las reformas.	Imagina que vives en un apartamento y tu vecino de al lado ha comenzado a hacer reformas en su casa. Las obras son ruidosas y están interfiriendo con tu vida diaria. No te dejan dormir, hacer las tareas de casa, escuchar la televisión y, en general, están afectando tu calidad de vida. Las reformas llevan varios días y ya estás cansado de escuchar el ruido. Decides ir al apartamento de tu vecino y contarle cómo el ruido te está afectando negativamente. • Le sugieres que limite los trabajos ruidosos a un par de horas. • Si no acepta limitar el ruido, llamarás a la policía. • Seguirás llamando a la policía cada vez que haga demasiado ruido. • Lo demandarás para que detengan las obras por completo. • También lo demandarás por las molestias que te está causando.

Tabla 3. *Guion para representar la situación 2 en el juego de rol*

Tarea de descripción de imágenes con fines fonéticos

Para la ejecución de esta última tarea lingüística, los informantes debían describir un total de dieciocho imágenes presentadas a través de la *tablet*. El propósito de esta tercera tarea es recoger material que permita analizar las diferencias acústicas entre hombres y mujeres, en concreto, para estudiar la divergencia en las pronunciaciones de las vocales y el fonema /s/ (Tabla 4), variable que en estudios previos como Brogan y Bolyanatz (2018) se relaciona con diferencias de género.

Vocal en la primera sílaba	**Posición del sonido /s/ en la palabra**		
	Inicial	**Media**	**Final**
/a/	*saco sapo*	*casa vaso*	*patos gatos*
/i/	*silla silo*	*misa risa*	*picos niños*
/o/	*soja/soya sopa*	*rosa sosa*	*toros fotos*

Tabla 4. *Palabras clave para la descripción de imágenes*

Los informantes desconocían la finalidad concreta del ejercicio, por tanto, y para asegurar la naturalidad del discurso, debían describir las imágenes mostradas en la Tabla 5 de manera detallada y no limitándose a leer un listado de palabras inconexas.

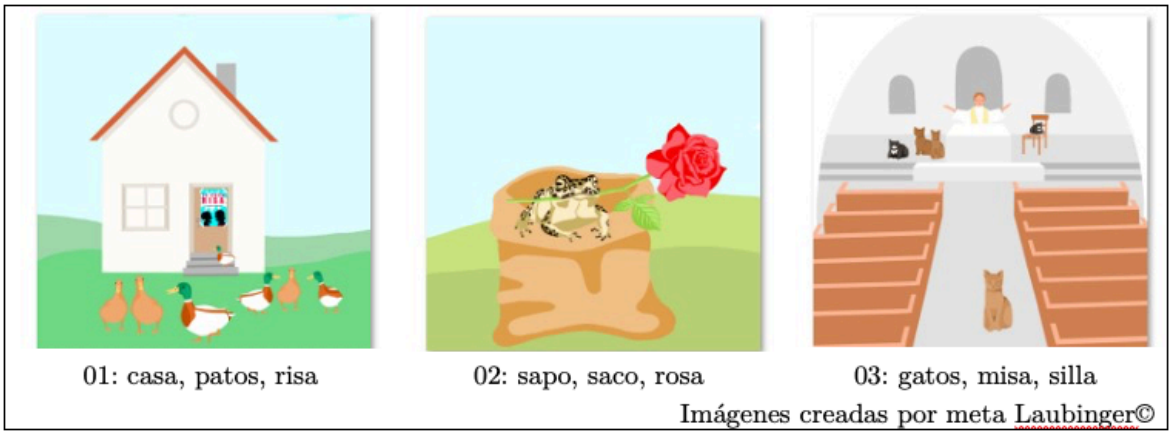

Tabla 5. *Muestra de las tres primeras imágenes de la tarea*

Esta tarea no se pudo realizar para el subcorpus CoLaGe-G-Diversity porque, este corpus se recogió de manera posterior, por tanto, los entrevistadores no dispusieron del mismo equipo técnico ni contaban con localizaciones fijas para efectuar las grabaciones. Las entrevistas se hicieron en lugares como cafeterías y espacios públicos en los que no se podía evitar la contaminación acústica.

Material no lingüístico: el cuestionario sociopsicológico

Para este proyecto se ha recogido un amplio conjunto de datos referidos a las características sociopsicológicas de los participantes atendiendo a diferentes parámetros, como son el autoritarismo de derechas, los estereotipos e ideologías de género, las actitudes hacia los homosexuales, el autoconcepto de género y la prototipicidad. Se describen, a continuación, los principales rasgos de cada uno de ellos.

Autoritarismo de derechas

Según Altemeyer (1996), este parámetro se define como un constructo multidimensional centrado en las facetas de sumisión, agresión y convencionalismo. Para el cuestionario realizado en el corpus CoLaGe, nos hemos servido de este valor para registrar las opiniones políticas de los informantes según una escala de 7 puntos (-3 = totalmente en desacuerdo, +3 = total-

mente de acuerdo). Algunas de las afirmaciones que debían valorar fueron, entre otras, «lo que nuestro país verdaderamente necesita, en lugar de más "derechos civiles", es una buena dosis firme de ley y orden» o «los días en que las mujeres eran sumisas pertenecen al pasado. El "lugar de la mujer" en la sociedad es donde ella quiera estar».

Estereotipos de género

Este segundo parámetro se relaciona con las expectativas sobre las características asociadas a mujeres y hombres (Hannover, 2006). Estas pueden darse en diferentes ámbitos, como por ejemplo, los rasgos que conforman la personalidad, las aficiones e intereses, así como los comportamientos o la apariencia (Deaux & Lewis, 1984; Deaux & LaFrance, 1998). Dado que el proyecto aborda la interconexión entre género y lenguaje, nos hemos centrado en la relación entre los estereotipos con dos dominios lingüísticos: la morfosintaxis y la voz. Además de relacionarlos con la personalidad como el dominio más destacado de los estereotipos de género (Rudman & Glick, 2008).

Para representar los estereotipos morfosintácticos de género se ha utilizado la encuesta MOGEST-s (Posio *et al.,* 2024). Este instrumento de análisis contiene doce pares de enunciados que difieren en el uso de rasgos morfosintácticos como la persona gramatical, las expresiones referenciales, el estilo (im-)personal, y el discurso directo, y se usa para comprobar si los informantes tienden a atribuir alguna de estas características a hablantes de diferentes géneros o de orientaciones sexuales. Así, en la tarea el informante debía marca cuál de las oraciones creía que había sido dicha por una mujer y cuál por un hombre.

Para medir los estereotipos de género con respecto a la voz, se seleccionaron doce características relacionadas con ella (por ejemplo, «fuerte», «suave» o «aguda») de una lista de cuarenta estereotipos explícitos del habla (Kachel *et al.,* 2018). Se pidió a los participantes que calificaran cada elemento en función de si tendían a asociarlo con mujeres u hombres en una escala de 7 puntos (1 = exclusivamente mujeres, 4 = igualmente mujeres y hombres, 7 = exclusivamente hombres).

Con respecto a la personalidad, se ha utilizado el *Inventario de roles sexuales* de Bem (Bem, 1974), creado originalmente como forma de evaluar el autoconcepto de rol de género. Este inventario mide los rasgos de personalidad esperables socialmente para las mujeres (por ejemplo, «afectuoso», «cálido») y para los hombres (por ejemplo, «asertivo», «dominante»). De la escala original, se seleccionaron nueve ítems de tipo femenino y se complementaron con tres nuevos («sumisa», «empática», «comprensiva»). Se pidió a los participantes que indicaran hasta qué punto consideraban que los ítems propuestos eran típicos de las mujeres o de los hombres en una escala de 7 puntos (1 = muy típico de las mujeres, 7 = muy típico de los hombres).

Ideologías de género

Para examinar las creencias sobre los roles de género aplicamos el *Cuestionario de orientación normativa sobre los roles de género* (Athenstaedt, 2000). En esta parte, los informantes indicaron su acuerdo con cada ítem en una escala de 7 puntos (-3 = totalmente en desacuerdo, +3 = totalmente de acuerdo) para ver medir de qué manera podían tener interiorizadas las normas relacionadas con los roles de género. Entre otras afirmaciones, se les presentaron algunas como, por ejemplo, «los hombres deberían tener permiso de paternidad con el nacimiento de sus hijos», «niños y niñas deberían realizar las mismas tareas de la casa» o «con el fin de

causar una buena primera impresión, es más importante que la mujer mantenga una buena apariencia que el hombre».

Actitudes hacia los homosexuales

Dado que la homonegatividad hacia los hombres suele ser más pronunciada que hacia las mujeres (así lo han demostrado, entre otros, Steffens & Wagner, 2005), nos hemos centramos en las actitudes hacia los hombres homosexuales.

Para la medición hemos utilizamos la escala SABA (Preuß *et al.*, 2020) basada en la propuesta de escenarios que recogen el componente afectivo y conductual de las posibles actitudes. La escala consta de cinco escenarios de la vida cotidiana (un ejemplo sería «usted tiene un hijo que va a la guardería, y se entera de que el profesor de su hijo ha salido recientemente del armario como gay). Se pidió a los participantes que indicaran su respuesta afectiva («¿Cómo de cómodo se siente en esta situación?») en una escala de 7 puntos (-3 = muy incómodo, +3 = muy cómodo). Para la respuesta conductual, se pidió a los participantes que calificaran su tendencia de aproximación/evitación específica del escenario (por ejemplo, «¿qué probabilidad hay de que intente cambiar la plaza de clase de su hijo?») en una escala de 7 puntos (-3 = muy improbable, +3 = muy probable).

Autoconcepto de género

El autoconcepto de rol de género podría definirse como la autoasignación de estereotipos de género a uno mismo, dando lugar a una autopercepción del grado de masculinidad o feminidad (Athenstaedt y Alfermann, 2011). Dado que el autoconcepto de rol de género se basa en estereotipos, también puede referirse a diferentes ámbitos como los rasgos de personalidad, intereses, comportamiento, apariencia, etc., contemplados anteriormente.

Para representar el dominio basado en la personalidad del autoconcepto de rol de género, seleccionamos las subescalas de feminidad y masculinidad positivas del *Cuestionario de Atributos Personales Ampliado* (Spence *et al.,* 1979). Mientras que la subescala de feminidad positiva contiene ocho ítems socialmente expresivos que son más deseables socialmente para las mujeres (por ejemplo, «emocional»), la subescala de masculinidad positiva abarca ocho ítems instrumentales que son más deseables socialmente para los hombres (por ejemplo, «independiente»). Todos los ítems se presentan como diferenciales semánticos de 5 puntos (por ejemplo, 1 = nada emocional; 5 = muy emocional.

Para medir el ámbito conductual del autoconcepto de género se utilizó la *Escala de Conducta de Género* (Athenstaedt, 1999). Contiene dos subescalas que abarcan conductas socialmente deseables de diversos ámbitos de la vida (por ejemplo, trabajo, hogar, ocio). De la subescala de conductas femeninas, que incluye conductas más típicas de las mujeres (por ejemplo, «dar un paseo por la ciudad»), se seleccionaron 28 ítems. De la subescala de comportamientos masculinos, que comprende comportamientos propios de los hombres (por ejemplo, «lavar el coche»), se eligieron 21 ítems. No se tuvieron en cuenta cuatro ítems de la escala original (es el caso de «quitar la nieve», que se excluyó porque difícilmente es una tarea que se haga en España o México). Todos los ítems debían valorarse en una escala de 7 puntos (-3 = completamente atípico, +3 = completamente típico).

Además de las representaciones específicas de cada ámbito, incluimos una medida transversal para el autoconcepto de género. La escala *Masculinidad y Feminidad Tradicional* (Kachel *et al.,* 2016) consta de seis ítems (por ejemplo, «Tradicionalmente, mis intereses se considerarían

como…») que se valoraron en una escala de 7 puntos (1 = muy masculino, 4 = igualmente masculino y femenino, 7 = muy femenino).

Prototipicidad

Además, recogimos datos sobre otro componente estructural, a saber, la prototipicidad para el propio grupo de género. Para ello, también utilizamos una medida pictórica (cf. Schubert & Otten, 2002). Esta vez, el círculo más pequeño representaba la propia visión del informante sobre sí mismo y el círculo más grande al grupo de hombres o mujeres en general. En cuanto a la centralidad, los círculos variaban en su distancia/proximidad entre sí a lo largo de siete imágenes (Tabla 6).

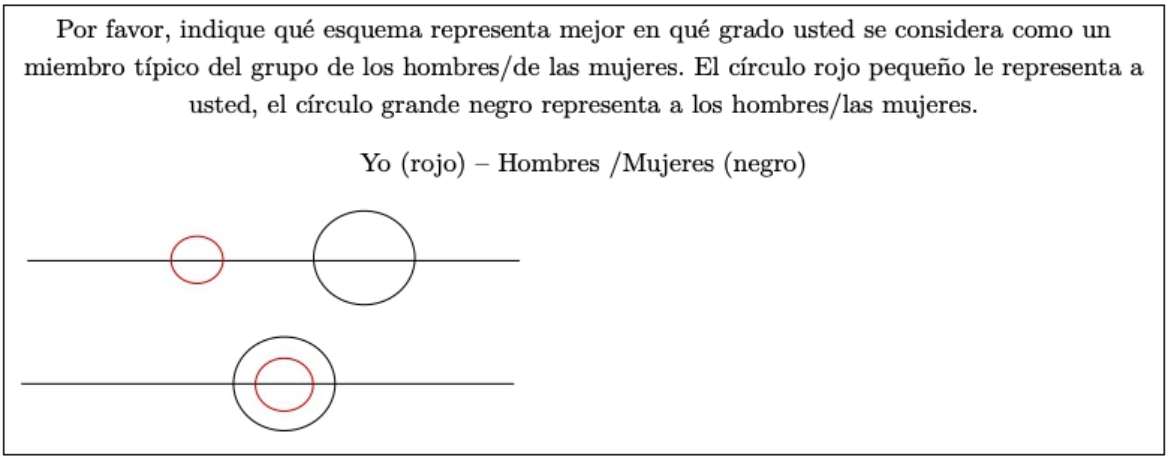

Tabla 6. *Extracto del cuestionario en el que se muestran los valores más extremos de prototipicidad*

3.2. Protocolo de tratamiento de los datos

Para la transcripción y codificación del corpus CoLaGe se ha utilizado un sistema de convenciones que combina dos modelos: por un lado, la propuesta de Jefferson (2004) para el análisis de la conversación, por otro, el sistema del grupo Val.Es.Co. (Briz *et al.*, 2002) para el estudio del español coloquial. Además, se han creado marcas de codificación propias.

La transcripción del material oral se ha realizado de manera alineada a través del programa ELAN Annotation (Max Planck Institute for Psycholinguistics, 2023). La ventaja de este *software* radica en que permite la creación de diferentes líneas de anotación para cada participante, así como para introducir observaciones y semiautomatizar la tarea posterior de pseudonimización del audio (ver apartado 3.2.3.).

Además, las marcas establecidas en el sistema de transcripción y codificación empleado se han configurado siguiendo el modelo de las normas TEI y del lenguaje de marcado XML que asegura la compatibilidad y legibilidad en ELAN.

En la Tabla 7 pueden observarse las convenciones de transcripción empleadas en el corpus CoLaGe:

Símbolo	**Descripción**	**Ejemplos**
[ ]	Solapamiento	A: ¿vienes ma[ñana]? B: [sí]
:	Alargamientos vocálicos y consonánticos	A: mañana: vamos a: la playa
PALABRA	Voz alta	A: es MUY importante
° palabra°	Susurro	A: es °muy importante°
( )	Transcripción ininteligible	A: no viene porque no ( ) esta semana
(palabra)	Transcripción dudosa	A: no viene porque no (ha cobrado) esta semana
-	Palabra cortada o reinicio	A: iremos a la pla- a la playa B: vamos a- iremos al mismo sitio de siempre
¡!	Exclamación	A: ¡qué fuerte!
¿?	Interrogación	A: ¿quién viene a la playa?
Mayúscula inicial	Se usa mayúscula en nombres propios	A: según *nombre1* ahora vive en *ciudad2*
<palabra>	Estilo directo	B: pues ayer me lo encontré y me dijo <¿sabes que me he mudado a ciudad2?>
>palabra<	Estilo indirecto	A: me dijo *nombre1* que > ahora vive en ciudad2<
/palabra/	Cambio de código o extranjerismo	A: bueno /xiqueta/ me voy a ir ya B: el viernes nos vamos de /shopping/
(RISAS)	Risas Cuando un informante habla entre risas se marca en la línea de observación de ELAN	A: (RISA) no me creo que Roberto se haya casado
(TOS)	Tos	A: estamos (TOS) todos esperándote
(SIC)	Indica que no se trata de un error de transcripción	A: hemos pedido cocretas (SIC) B: me se (SIC) ha roto el pantalón
ONU	Las siglas y acrónimos se transcriben en forma abreviada, no como se pronuncia	A: me pidieron el DNI para servirme una copa de vino

Otras indicaciones:

· Las pausas no se marcan en la transcripción, ya que son apreciables en ELAN, con la segmentación en grupos entonativos.

· Los números se transcriben en letra (54, por tanto, se transcribiría como cincuenta y cuatro).

· No se utilizan signos de puntuación propios del escrito, a excepción de los signos de interrogación y exclamación.

· Se identifica a los hablantes con las letras E (entrevistador/a) e I (informante).

Tabla 7. *Sistema de transcripción utilizado en el corpus CoLaGe*

Pseudonimización de los datos e identificación de los archivos

Para proteger la identidad de los hablantes, en el corpus CoLaGe se ha implementado un sistema de pseudonimización[6] de los datos personales en el que todos los identificadores directos, esto es, nombres propios, nombres de lugares, etc., que pudieran permitir la identificación del informante, han sido sustituidos por otros identificadores genéricos del tipo *nombre1, lugar1, empresa1, plaza1*. Además, el fragmento de audio correspondiente ha sido borrado de manera semiautomática utilizando de manera combinada el *software* de anotación ELAN y el editor de audio Audacity.

Asimismo, para la identificación de los archivos se ha implementado un código individual que ofrece la información sociolingüística principal de cada uno de los participantes, esto es, los metadatos (Tabla 8), sin utilizar elementos que puedan llevar a su identificación. De esta manera, encontramos los metadatos en el propio código identificador del archivo.

Código	**Significado**
E	Entrevista
F	Tarea fonética
R	Juego de rol
VLC	Valencia
GDL	Guadalajara
GDL-D	Guadalajara Diversity
M	Mujer
H	Hombre
NB	No binario

[6]Empleamos el término *pseudonimización* en lugar de *anonimización*, ya que según *el Language Bank of Finland*, repositorio en el que se alojará el corpus, el hecho de poder escuchar la voz del informante hace que nos encontremos ante un identificador potencial. Por eso, no se puede hablar de *anonimización* propiamente dicha.

HET	Heterosexual
LES	Lesbiana
BIS	Bisexual
PAN	Pansexual
GAY	Gay
OTH	Otros
30	Grupo 1, 30-40 años
60	Grupo 2, 60-70 años
01 y siguientes	Indica el orden de participación del informante en la recopilación del corpus

Ejemplo: R_VLC_M_HET_30_05
En este caso nos encontraríamos ante un archivo que contiene la tarea juego de rol en la que participa una mujer de Valencia, heterosexual, del grupo de edad 1 y que fue la 5.ª persona en participar en el estudio.

Tabla 8. *Código de identificación de los archivos*

4. Consideraciones finales

Con este nuevo corpus oral del español, de más de un millón de palabras, se amplía la nómina de corpus orales del español disponibles en línea y se ofrece una herramienta valiosa para la comunidad científica, concretamente, para el análisis lingüístico, sociolingüístico y sociopsicológico comparado entre dos variedades del español.

El corpus estará disponible para su consulta y descarga en el repositorio *The Language Bank of Finland*[7], coordinado por el consorcio nacional FIN-CLARIN. En este portal, cualquier usuario podrá descargar al material lingüístico debidamente pseudonimizado y, en el caso de ser necesario, acceder a las grabaciones de audio, también carentes de identificadores directos, previa solicitud de acceso. Los audios de cada una de las tareas se encuentran en formato .wav y van acompañados de la transcripción en ELAN en el caso de la entrevista sociolingüística y el juego de rol; para la tarea de descripción de imágenes con fines fonéticos el audio tiene asociado un archivo de PRAAT (TextGrid). Los materiales se completan con las transcripciones en formato .xlxs, que permiten trabajar con el corpus sin necesidad de manejar el *software* específico.

Cabe subrayar el valor añadido que este corpus ofrece al estudiar las interrelaciones entre el sexo y el género de los hablantes, los roles sociales y las expectativas de género en la sociedad, utilizando datos sociolingüísticos y de la psicología social. Su novedad radica en que combina la variable *sexo* con la variable *género*, a través de las diferentes escalas de

[7] https://www.kielipankki.fi/corpora/

autoconcepto recogidas, descritas en este trabajo; mientras que en la mayoría de los corpus orales del español solamente encontramos la estratificación binaria de sexo[8] (hombre/mujer).

5. Bibliografía

Albelda Marco, Marta (2022): «Los corpus del español hablado y los estudios pragmáticos». En Giovanni Parodi, Pascual Cantos-Gómez y Chad Howe (eds.), The Routledge Handbook of Spanish Corpus Linguistics, 223-238, Routledge. https://doi.org/10.4324/9780429329296

Altemeyer, Bob (1996): *The authoritarian specter.* Harvard University Press.

Athenstaedt, Ursula (1999): «Geschlechtsrollenidentität als mehrfaktorielles Konzept. Ein kritischer Beitrag zur Androgynieforschung». En U. Bock & D. Alfermann (Hrsg.), *Querelles. Jahrbuch für Frauenforschung,* 183–199. Metzler.

Athenstaedt, Ursula (2000): «Normative Geschlechtsrollenorientierung: Entwicklung und Validierung eines Fragebogens». *Zeitschrift für Differentielle und Diagnostische Psychologie, 21*(1), 91–104. https://doi.org/10.1024//0170-1789.21.1.91

Athenstaedt, Ursula & Alfermann, Dorothee (2011): *Geschlechterrollen und ihre Folgen.* Kohlhammer.

Bem, Sandra L. (1974): «The measurement of psychological androgyny». *Journal of Consulting and Clinical Psychology, 42*(2), 155–162. https://doi.org/10.1037/h0036215

Briz Gómez, Antonio & Grupo Val.Es.Co. (2002): *Corpus de conversaciones coloquiales.* Anejos Oralia. Arco Libros.

Briz Gómez, Antonio y Samper Hernández, Marta (2022): «Estudio de variación situacional en corpus orales del español», en Giovanni Parodi y otros, (eds.), *Lingüística de corpus en español / The Routledge Handbook of Spanish Corpus Linguistics*, Routledge, 309-324. https://doi.org/10.4324/9780429329296-24

Brogan, Franny D., & Bolyanatz, Mariska A. (2018): «A sociophonetic account of onset /s/ weakening in Salvadoran Spanish: Instrumental and segmental analyses». *Language Variation and Change*, 30(2), 203–230.

Carcelén Guerrero, Andrea (2024): «Novedades en los corpus digitales para el estudio del español oral». *Normas*, 14(1), 241-260. https://doi.org/10.7203/Normas.v14i1.29929

Comisión Europea (2017): Eurobarómetro especial 465: Igualdad de género 2017. European Commission, Directorate-General for Communication. http://data.europa.eu/88u/dataset/S2154_87_4_465_ENG

Deaux, Kay & Lewis, Laurie L. (1984): «Structure of gender stereotypes: Interrelationships among components and gender label». *Journal of Personality and Social Psychology, 46*(5), 991–1004. https://doi.org/10.1037/0022-3514.46.5.991

[8] Véase, por ejemplo, el Corpus del Español del siglo XXI (CORPES XXI), el corpus Ameresco (América y España Español Coloquial), el Corpus Val.Es.Co. (Valencia Español Coloquial), el corpus ESLORA o el Corpus Oral y Sonoro del Español Rural (COSER).

Deaux, Kay & Lafrance, Marianne (1998): «Gender». En Daniel T. Gilbert, Susan T. Fiske, & Gadner Lindzey (eds.), *The handbook of social psychology*, 788–827. McGraw-Hill.

ELAN (Version 6.9) [Computer software]. (2024): Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Hausmann, Ricardo *et alii* (2014): *The Global Gender Gap Report 2014*. World Economic Forum, Geneva.

INEGI (2016): *Encuesta Nacional sobre la Dinámica de las Relaciones en los Hogares* (ENDIREH). Instituto Nacional de Estadística y Geografía, México. https://www.inegi.org.mx/programas/endireh/2016/

Jefferson, Gail (2004): «Glossary of transcript symbols with an introduction». En Gene Lerner (ed.), *Conversation Analysis: studies from the first generation.* John Benjamins, 13-31.

Kachel, Sven *et alii* (2016): «Traditional masculinity and femininity: Validation of a new scale assessing gender roles». *Frontiers in Psychology,* 7(956). https://doi.org/10.3389/fpsyg.2016.00956

Kachel, Sven *et alii* (2018): «Do I sound straight? – Acoustic correlates of actual and perceived sexual orientation and masculinity/femininity in men's speech». *Journal of Speech Language and Hearing Research*, *61*(7), 1560–1578. https://doi.org/10.1044/2018_JSLHR-S-17-0125

Leiner, Dominik J. (2019): SoSci Survey (Version 3.1.06) [Computer software]. Disponible en https://www.soscisurvey.de

Llisterri, Joaquim (2021): «Corpus para investigar sobre el componente fónico en español como LE/L2». En Mar Cruz y Javier Muñoz (eds.), *e-Research y español LE/L2.: investigar en la era digital*, 164-196, Routledge.

Posio, Pekka *et alii* (2024): «Morphosyntactic stereotypes of speakers with different genders and sexual orientations: an experimental investigation». *Linguistics*, 62(6), 1581-1625. https://doi.org/10.1515/ling-2022-0143

Preuss, Sabine *et alii* (2020): «Using scenarios for measuring the affective and behavioral components of attitudes toward lesbians and gay men: Validation of the SABA scale». *Archives of Sexual Behavior, 49*(5), 1645–1669. https://doi.org/10.1007/s10508-020-01653-7

Prince, V. (1985): «Sex, gender, and semantics». The Journal of Sex Research, 21, 92–96.

Recalde, Montserrat y Vázquez, Victoria (2009): «Problemas metodológicos en la formación de corpus orales». En Pascual Cantos y Aquilino Sánchez (eds.): *A Survey on Corpus-based Research,* AELINCO, 37-49.

Reglamento (EU) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (*Reglamento General de Protección de Datos*).

Schubert, Thomas W. & Otten, Sabine (2002): «Overlap of self, ingroup, and outgroup: Pictorial measures of self-categorization». *Self and Identity, 1*(4), 353–376.

https://doi.org/10.1080/152988602760328012

Spence, Janet T. *et alii* (1979): «Negative and positive components of psychological masculinity and femininity and their relationships to self-reports of neurotic and acting out behaviors». *Journal of Personality and Social Psychology*, *37*(10), 1673–1682.

Social Watch (2012): «Gender Equity Index». En *Sustainable Development: The Right to a Future*. http://www.socialwatch.org/report2012.

Rudman, Laurie A. & Glick, Peter (2008): *The social psychology of gender*: *How power and intimacy shape gender relations.* Guilford Press.

Steffens, Melanie & Wagner, Christof (2005): «Attitudes toward lesbians, gay men, bisexual women, and bisexual men in Germany». *Journal of Sex Research*, 41, 137–149. https://doi.org/10.1080/00224 4904095522 22

Thompson, Edward H. & Pleck, Joseph H. (1986). «The structure of male role norms». *American Behavioral Scientist,* 29(5), 531-543.

West, Candance & Zimmerman, Don H. (1987): «Doing gender». *Gender & Society* 1(2), 125–151.

Uclés Ramada, Gloria *et alii* (en preparación): *Corpus for the study of Language and Gender in Mexico and Spain* (CoLaGe).