

**En los orígenes del estudio de la variación lingüística. Interpretación optimista de los resultados de un estudio cuantitativo de variación léxica en la ciudad de Málaga**

## **Introducción**

Las lenguas manifiestan múltiples posibilidades de variación que se observan en todos los niveles de análisis lingüístico. El efecto de los marcadores de hablante se deja sentir en todos los componentes del lenguaje: desde la fonética hasta la pragmática, pasando por el léxico o la gramática. Quizás, uno de los ámbitos donde más se aprecien las consecuencias de esa influencia sea el léxico. Empleamos palabras que usan o usaban nuestros padres; las palabras nos unen a nuestros vínculos más fuertes y aprietan lazos de solidaridad con aquellos que forman grupos con los que nos identificamos (Childs y Mallinson, 2006).

En este trabajo utilizamos el estudio de la variable léxica para insistir en el hecho demostrado de que los hablantes estamos condicionados, desde el punto de vista lingüístico, por nuestra identidad social. Cuanto más parecidas son las características sociales de dos sujetos, más probabilidades existirán de que la naturaleza de sus producciones léxicas sean también similares; por el contrario, la variación de los rasgos sociales será un factor determinante para que en sus realizaciones léxicas se observe diferencia.

En el ya amplio panorama histórico de la variación lingüística, la investigación sobre la variación léxica ha sido limitada debido, esencialmente, a las dificultades metodológicas que acarrea su estudio y a la problemática que supone establecer equivalencias semánticas entre variantes aparentemente sinónimas. La solución adoptada en este trabajo para solventar estos inconvenientes ha sido, en primer lugar, el uso de la metodología propia de los estudios de disponibilidad léxica para la recogida masiva de material de forma sencilla. En segundo lugar, realizamos un planteamiento teórico basado en la asociación de los conceptos de ‘disponibilidad’ y ‘prototipicidad’ para obtener

valores estadísticos originales que han demostrado su eficacia para medir la centralización/descentralización léxica de individuos que pertenecen a grupos sociales (Ávila y Villena, 2010). El empleo del asistente semiautomático de análisis de corpus *Annotate* permite obtener información precisa de cada uno de los términos contenidos en las listas estudiadas. De esta manera, los inconvenientes derivados de los valores connotados o particulares del uso idiolectal de las palabras quedan minimizados.

## **1. Planteamiento**

### **I**

La posición de los hablantes en la estructura social y en las redes de relaciones determina, en buena medida, su conducta lingüística. La pertenencia a las clases sociales, la etnia, el género o la edad, por un lado, y la fuerza y densidad de los vínculos en el interior de la red social, por otro, son factores que pueden predecir el comportamiento lingüístico de los sujetos. No obstante, debemos interpretar lo anterior con cautela pues, de hecho, parte del comportamiento lingüístico individual puede seguir siendo variable a pesar de que se den condiciones de semejanza estructural y reticular. Para tratar este último reducto de la variación lingüística se suele plantear un modelo que considere la historia social de los individuos. A partir de la teoría de la red social, es posible estudiar las motivaciones individuales que conducen a los hablantes a reafirmarse en los patrones propios o a adscribirse a otros diferentes (Labov y Harris, 1986; Milroy y Milroy, 1992).

En la construcción de una teoría variacionista completa es necesario, pues, organizar el trabajo en tres niveles diferentes fuertemente interrelacionados (Villena y Ávila 2014; Villena, Ávila y von Essen, 2017):

1. Nivel estratificacional
2. Nivel reticular
3. Nivel individual

Esos niveles son el objeto de otras tantas teorías (de la estratificación, de la red social y de la historia social). En cada una de ellas se explica el grado específico de semejanza en la conducta lingüística de los hablantes (estratificacional, reticular, intencional), tal y como se refleja en la Tabla 1.

Tabla 1. Niveles, teorías y clases de semejanza en el uso lingüístico

Nivel	Teoría	Semejanza
Estructura social	Estratificación social	Estructural
Estructura intermedia	Red social	Reticular
Individuo	Historia social	Intencionada

La relación entre los tres niveles considerados es esencial en el estudio de la variación. La posición de los hablantes en la estructura social condiciona primariamente su comportamiento lingüístico por cuanto el contexto de adquisición de la variedad materna es el familiar; si, además, esta situación se combina con una estructura reticular densa y un tipo comunitario cerrado con pocos contactos externos, se refuerzan valores de solidaridad (frente a valores de estatus) y se acrecienta la tendencia vernacular; por último, la situación del individuo en la red social, el tipo de vínculos que lo une con sus miembros y la biografía social que lo identifica y lo define con respecto a sus pares reticulares puede reforzar o modificar la tendencia o el consenso de la comunidad local, aunque este último nivel es más una opción intencional (voluntaria) del hablante.

## II

La variación léxica introduce cierta controversia en el concepto tradicional de ‘variable sociolingüística’. A veces, resulta difícil identificar formas equivalentes que expresen el

mismo significado semántico. A menudo existen pocos medios para acceder a la manera en que los hablantes seleccionan las palabras con lo que pueden existir desajustes entre la intencionalidad del hablante y la interpretación de los destinatarios e investigadores (Geeraerts, 2010: 823-824.) Para intentar minorizar estos efectos adversos, Geeraerts propone un modelo explicativo de la variación léxica basado en la siguiente tipología:

1. La variación formal onomasiológica (*formal onomasiological variation*, FOV); el hablante elige una palabra entre otras que expresan el mismo concepto: *macilento-canijo* para expresar el significado ‘delgado’.
2. La variación conceptual onomasiológica (*conceptual onomasiological variation*, COV); el hablante selecciona temas o asuntos sobre los que hablar, no las palabras que usará: por ejemplo, referirse a temas técnicos en vez de a asuntos relacionados con las distracciones.
3. La variación relacionada con la situación y con las propiedades sociales del hablante (*speaker and situational related variation*, SSV); el individuo se verá afectado por la interacción del contexto comunicativo y, sobre todo, por otras características relativamente estables, como su identidad social.

Como es lógico, las imbricaciones entre estas dimensiones generales son continuas: por ejemplo, la elección entre *canijo-macilento* para expresar el concepto ‘delgado’ es un caso de una interacción SSV-FOV debido a que la preferencia por un término más o menos formal vendrá dada por la formalidad de la situación de habla. Sin embargo, la elección de la expresión +formal *macilento* también podría estar motivada por factores temáticos y de educación formal de los hablantes, en cuyo caso nos encontraríamos con una interacción COV-FOV-SSV (Geeraerts 2010: 823).

Nosotros vamos a prestar atención a los tres tipos de variación, pues la metodología empleada —propia de los estudios de disponibilidad léxica— introduce a los informantes

el estímulo temático sobre el que va a actualizar un listado de palabras; después, los sujetos seleccionan el vocablo que, según su criterio, mejor representa al conjunto de elementos que forman el estímulo propuesto, condicionado tanto por la situación concreta en la que se desarrolla la prueba como por sus identidades sociales. En nuestro análisis final prestaremos atención a cómo afectan las características sociales estructurales a la selección léxica y, dentro de ella, al tipo o la naturaleza del léxico seleccionado por los grupos de iguales.

## **2. Hipótesis**

Las hipótesis de partida son las siguientes:

1. Existe una coincidencia tanto en la naturaleza estructural como en la amplitud del léxico compartido por los hablantes que pertenecen y se identifican con grupos sociales amplios. Aspectos como el género, la posición social, el nivel educacional o la edad son factores que determinan la preferencia de los grupos de iguales para actualizar en sus discursos determinado tipo de léxico: técnico, dialectal o extranjero, por ejemplo. Factores como la amplitud léxica estarían también correlacionados con los factores sociales señalados. Geeraerts, Grondelaers y Bakema (1994: 7) consideran que en el proceso de selección onomaseológica de una palabra para expresar un referente existen dos aspectos diferenciados:
  - a. La elección de una categoría conceptual para identificar o definir el referente, y
  - b. La elección de un ítem léxico dentro de esa categoría.

Determinados individuos afianzan su identidad social a través de la elección de las categorías conceptuales señalada en el apartado a), entendiendo por ‘categoría conceptual’ aspectos estructurales del léxico como su especificidad terminológica. Esta elección consciente puede hacer que determinados sujetos actualicen categorías conceptuales que los identifique como grupo frente al resto de la sociedad.

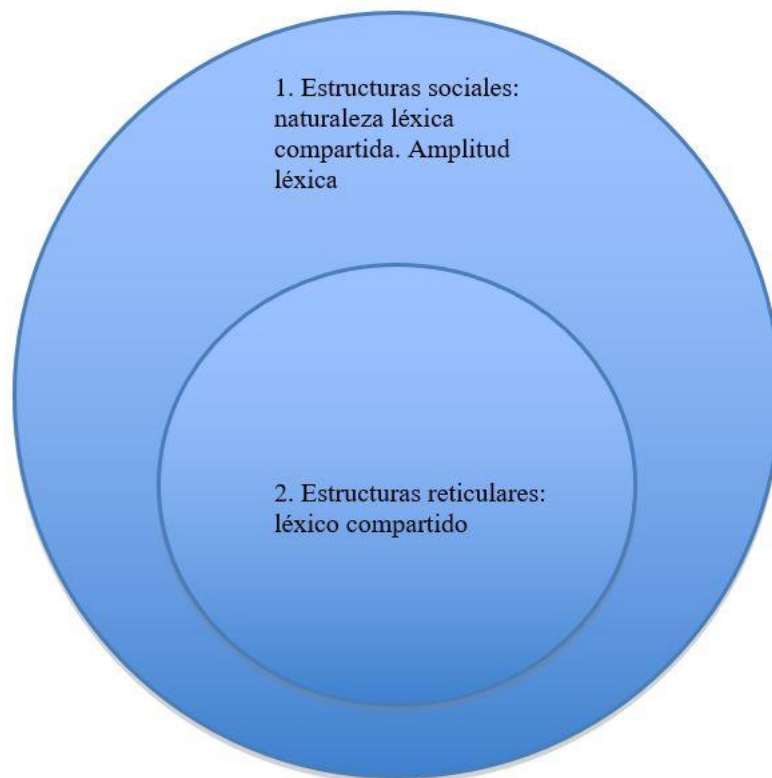
2. En las estructuras intermedias, cobra más importancia la elección del ítem léxico específico (apartado b). En las redes sociales formadas por individuos que interactúan diariamente, existe una coincidencia no solo en aspectos de la estructura gramatical o sustancial del léxico sino, sobre todo, en las unidades léxicas empleadas.
  - a. Las redes sociales pueden estar sostenidas por el léxico que comparten sus miembros. A medida que las redes son más densas y los lazos más estrechos y múltiples, el léxico compartido señala marca grupal.
  - b. Los miembros de redes densas de nudos apretados manifiestan su identidad mediante la divergencia léxica respecto al resto de grupos. A veces, esa divergencia puede estar basada en el empleo de léxico tradicionalmente considerado vernacular; otras veces, las diferencias pueden marcarse a través del empleo de léxico jergal, vulgar, elitista, etc. Las redes sociales más densas y compactas muestran mayor reticencia al abandono de las formas léxicas tradicionales y a la convergencia léxica con el estándar socialmente aceptado —al menos, en los registros más informales—. Así, la dicotomía forma tradicional (particular, geográficamente restringida con prestigio encubierto)-forma estándar (general, socialmente aceptada con prestigio manifiesto) puede entenderse

como un refuerzo de los lazos: la identificación con la forma tradicional funciona como una fuerza conservadora que resiste las presiones en forma de cambios externos a la red. Los sujetos de estas redes mantienen rasgos vernaculares y rechazan usos estándares considerados legítimos en la comunidad general (Ávila, 2017).

En las redes poco densas, de nudos laxos y simples, la tendencia podría ser la contraria: existe un proceso de convergencia con el léxico estándar y un abandono relativamente consciente del léxico autóctono.

Estaríamos, en consecuencia, ante espacios de identificación grupal que utiliza al léxico como elementos de integración de sus miembros. El Dibujo 1 esquematiza la situación:

Dibujo 1. Espacios de identificación grupal basados en la variación léxica



Si estas hipótesis son ciertas, por un lado, los individuos que pertenecen a determinados estratos dentro de grupos estructurales previamente establecidos (clase, género, generación) emplearían el mismo tipo de léxico frente a otros sujetos de estratos diferentes. Como es lógico, estos individuos también compartirán parte de su léxico, pero el mayor identificador grupal estará determinado por su naturaleza, estructura y amplitud; por otro lado, los miembros de grupos reticulares densos y múltiples encontrarán ese rasgo de identidad grupal más en el léxico compartido que en su categoría conceptual, estructura o amplitud.

### **3. Objetivos**

Dada la complejidad de la tarea, en el presente estudio pretendemos demostrar, únicamente, parte de las hipótesis anteriores. Queremos comprobar que, en efecto, la posición de los hablantes y sus características personales son factores que predicen cierto grado de variación que se manifiesta en su diferente amplitud léxica y en el empleo de léxico especializado.

Para conseguir este objetivo principal, se plantean algunos objetivos asociados:

- a) obtener un corpus léxico de tamaño suficiente para corroborar nuestras hipótesis;
- b) estudiar una población diversa que muestre variación suficiente para contribuir a una mejor comprensión de los mecanismos de la interacción sociolingüística en la comunidad estudiada;
- c) crear una herramienta que nos permita analizar eficazmente las características del léxico obtenido.

## **4. Método**

### *4.1. Muestra*

Usaremos el mismo corpus léxico que posibilitó el estudio del vocabulario disponible de los hablantes de la ciudad de Málaga (*Proyecto CONSOLEX*). Es importante considerar la situación del español de Málaga en el ámbito del español de España. En el contexto del español de Europa, los patrones estándares de la lengua española están contruidos históricamente y socialmente normativizados a partir de las características del español de

Castilla (Castilla y León). Se trata de una variedad lingüística relativamente alejada de la que se emplea en Málaga, sobre todo si se consideran patrones de actuación fonéticos.

El objetivo del *Proyecto CONSOLEX* fue producir un diccionario de disponibilidad léxica. Su originalidad consistió en iniciar una tradición de estudios sobre la variación social del léxico disponible en la ciudad de Málaga (Ávila y Villena, 2010) que continúa en la actualidad. Se usó una muestra representativa con afijación uniforme de hablantes estratificada según tres variables sociales: sexo, edad y nivel de educación formal (N=72). En el presente estudio se ha analizado con *Annotate* el léxico contenido en las listas de disponibilidad de 36 informantes seleccionados siguiendo la distribución que aparece en la Tabla 2 referida al total de la muestra.

Tabla 2. Distribución de la muestra de hablantes del estudio CONSOLEX

Hombres					Mujeres				
Educación	0	1	2	Total	Educación	0	1	2	Total
Edad					Edad				
20-34	4	4	4	12	20-34	4	4	4	12
35-54	4	4	4	12	35-54	4	4	4	12
>55	4	4	4	12	>55	4	4	4	12
Total	12	12	12	36	Total	12	12	12	36

Esta distribución es idéntica a la empleada en el *Proyecto de Estudio del Español de España y América* (Proyecto de Investigación PRESEEA) coordinado por el profesor Francisco Moreno Fernández. Sus principios metodológicos están basados en la evidencia de que una muestra estratificada por cuotas es productiva: por un lado, favorece la selección de los informantes y, por otro, facilita la comparación estadística dentro de la muestra y entre muestras diferentes.

Para estructurar la investigación sobre la disponibilidad léxica en la ciudad de Málaga se asumieron los fundamentos metodológicos del *Proyecto Panhispánico sobre Disponibilidad Léxica* (PPHDL) coordinado por el profesor Humberto López Morales.

Por un lado, para la obtención de las listas de léxico se emplearon 20 campos de experiencia compartida o centros de interés. De ellos, 16 eran tradicionalmente empleados en los estudios enmarcados en el PPHDL (01. El cuerpo humano; 02. La ropa; 03. Partes de la casa (sin los muebles); 04. Los muebles de la casa; 05. Alimentos; 06. Objetos colocados en la mesa para la comida; 07. La cocina y sus utensilios; 08. La escuela: muebles y materiales; 09. Calefacción, iluminación y medios de airear un recinto; 10. La ciudad; 11. El campo; 12. Medios de transporte; 13. Trabajos del campo y del jardín; 14. Los animales; 15. Juegos y distracciones; 16. Profesiones y oficios) y 4 fueron añadidos por el interés particular que tenían para el estudio en Málaga (17. Dinero y economía; 18. Internet y ordenadores; 19. Los colores; 20. El mar). Cada informante dispuso de 2 minutos para elaborar unas listas que posteriormente eran editadas siguiendo las normas propuestas por el PPHDL. Por otro lado, para obtener datos sociológicos y reticulares de los informantes, cada uno rellenaba un cuestionario socio-reticular y de actitudes lingüísticas compuesto por 47 preguntas que era almacenado en la base de datos del *Proyecto CONSOLEX*.

Los beneficios de trabajar bajo el auspicio de estos dos proyectos de referencia fueron innumerables. En relación con el presente estudio, la principal ventaja reside en que todas las voces analizadas están conectadas con el informante que las actualizó y con su centro de interés. Esto permitió establecer múltiples relaciones entre las características léxicas y morfológicas de las voces estudiadas y las variables de hablante utilizadas en la investigación, con lo cual las posibilidades de la herramienta *Annotate* se enriquecieron considerablemente.

#### *4.2. Estudios sobre el léxico disponible*

La idea que subyace a los estudios de disponibilidad léxica es averiguar qué palabras se actualizan antes y con mayor frecuencia en un número reducido pero representativo de campos de experiencia. Estas investigaciones tienen una larga trayectoria plasmada en la edición de diccionarios del léxico disponible correspondiente a distintas sintopías en todo el mundo. Estos diccionarios permiten aplicaciones en diferentes ámbitos, especialmente en la enseñanza de la lengua materna o de lenguas extranjeras (Ávila, 2016, 2019; Trigo, Romero y Santos 2019), pero también en dominios como la psicolingüística o la lexicografía (Ávila y Sánchez, 2010: 42-45).

Habitualmente, los procedimientos de recogida del material del léxico disponible son sencillos y permiten obtener corpus fiables y de tamaño suficiente en un tiempo razonable. Los participantes suelen ser colectivos de informantes previamente establecidos —generalmente grupos de escolares—. El material se obtiene usando pruebas asociativas consistentes en la propuesta de centros de interés o campos de experiencia compartida —como hemos señalado, casi siempre los mismos, con lo que las comparaciones entre los diferentes trabajos suelen ser frecuentes—. Los informantes disponen de un tiempo limitado para generar sus listas y, posteriormente, su contenido es editado según criterios de lematización previamente consensuados. A partir de ese momento, sin embargo, el investigador deja de obtener información relevante sobre las formas que componen las listas, a no ser que recurra a sus anotaciones de campo o a la engorrosa tarea que supone recuperar los listados originales en formato papel.

Aunque este tipo de estudios centrados en la estructura léxica de la lengua (esto es, en las palabras) resultan por sí mismos de notable interés, es paradójico que no se haya aprovechado en mayor medida la información léxica que contiene las listas y se limiten, muchas veces, a constatar la más que evidente aparición de sustantivos frente a otras

categorías gramaticales o a realizar estudios sobre los errores ortográficos que contienen (Trigo, Romero y Santos, 2019).

Hace más de veinte años que en la Universidad de Málaga tomamos contacto con los procedimientos y rutinas de análisis del léxico disponible. La misión nos pareció metodológicamente interesante y muy productiva. Cuando nos incorporamos a la tarea ya existía una red de estudiosos con una larga trayectoria en este ámbito que formaba una sólida estructura de investigación. Ello nos permitió adaptarnos sin dificultad a un esquema de trabajo perfectamente asentado de resultados constantes. En aquel momento existían protocolos de actuación que abarcaban todas las fases del proceso que supone la obtención de los listados de léxico disponible: desde el diseño de la muestra hasta la explotación de los resultados, pasando por la metodología de recogida de datos, el procesamiento y lematización de las palabras y la obtención de los listados de disponibilidad. Además, esta consolidada red de investigadores había desarrollado programas informáticos capaces de generar automáticamente el índice de disponibilidad léxica (LexiDisp) y fomentaba la cooperación y el intercambio de datos mediante plataformas electrónicas que utilizaban internet como un medio eficaz de gestión y organización (DispoLex). Todo ello, en fin, bajo la coordinación y tutela de experimentados investigadores supervisados por el profesor Humberto López Morales, empeñado en la labor de obtener un Diccionario Panhispánico de Disponibilidad Léxica.

Sin embargo, desde la publicación de nuestros primeros trabajos en este ámbito (Ávila, 2006), notamos que, a pesar de su tradición y arraigo, los estudios sobre disponibilidad léxica adolecían de una teoría general que abarcase todas sus dimensiones y respondiese a los interrogantes esenciales que, a veces, han desmerecido el esfuerzo que supone elaborar un listado de estas características. Bien sea por la ausencia de una explicación teórica sólida del fundamento léxico-estadístico o cognitivo general que

subyace a estos estudios (Ávila y Sánchez, 2010: 37-81), bien por las autolimitaciones que la metodología de recogida y análisis de datos impone a la explotación de las listas de disponibilidad léxica (Sánchez-Sauss Laserna, 2020), lo cierto es que sus posibilidades se han visto frecuentemente mermadas. A veces, estas limitaciones han sido superadas desde planteamientos novedosos como, por ejemplo, focalizar el interés más en los hablantes que en las palabras de las listas gracias a la construcción de modelos de centralización y descentralización léxica capaces de caracterizar la posición de los individuos con respecto al resto de la población (Ávila y Sánchez, 2010, Ávila y Villena, 2010: 177-280). Ávila y Sánchez (2010) usaron la teoría de conjuntos difusos para determinar la compatibilidad de cada vocablo encontrado en los listados de disponibilidad léxica con el centro de interés en el que aparece. Gracias a esa compatibilidad se realizaron cálculos sobre la capacidad léxica individual: el Índice de Descentralización (IDL) consideraba “tanto el número de elementos léxicos enunciados como su relevancia en el centro de interés” (2010: 69). Se otorgó importancia a la presencia en las listas de vocablos que se alejaban del concepto central (o prototipo), pues su presencia es indicativo de una capacidad léxica mayor: “son vocablos de uso más restringido y, por tanto, se consideran de más difícil acceso para el conjunto de los hablantes” (2010: 69). Esto quiere decir que, junto al vocabulario prototípico de cada campo, también existe otro léxico más especializado o más preciso. La aparición de este léxico menos prototípico sugiere que el hablante que lo actualiza posee una capacidad léxica individual que se aleja del dominio común. El modelo propuesto permitió obtener el grado de compatibilidad —entre 0 y 1— de cada vocablo con la categoría conceptual representada por cada centro de interés y, a partir de él, establecer también el grado de especificidad de cada uno de ellos (poca compatibilidad supondrá que no es central o prototípico, sino específico y poco habitual en el uso de la comunidad). De la misma manera, este planteamiento

permite calcular la descentralidad léxica de un individuo, es decir, la capacidad que tiene para acceder al léxico no común o especializado (2010: 79).

En este trabajo nos hemos beneficiado tanto de los avances propuestos en el trabajo mencionado de Ávila y Sánchez (2010) como de las ventajas proporcionadas por una herramienta de análisis lingüístico capaz de obtener un mayor rendimiento de las listas de léxico disponible (*Annotate*). Empeñados en aprovechar toda la información que estas listas nos ofrecen, creamos esta herramienta de análisis semiautomática que abarca diferentes niveles: gramatical, morfológico, ortográfico, de vigencia dialectal, técnica, de uso o registro que, en definitiva, nos permite obtener información de cada una de las formas recogidas en las listas de disponibilidad. A partir de esta información, las posibilidades de investigación se disparan hacia límites hasta hace poco inéditos en este tipo de estudios sobre el léxico. En concreto, aquí explotaremos la información procedente de la base de datos *Annotate* referida al grado de especificidad técnica de los elementos que contiene.

### *4.3. Annotate: una herramienta semiautomática para el análisis del léxico*

#### 4.3.1. Justificación

En las investigaciones lingüísticas es habitual el etiquetado de corpus con el objetivo de añadir información que permita posteriores análisis y evite interpretaciones ambiguas de los datos. Se suelen considerar dos planteamientos para afrontar el problema de etiquetar contenidos léxicos:

1. Etiquetación automática: un algoritmo determina la marcación de los términos de forma autónoma a partir de un conjunto de reglas o modelos previamente establecido.

2. Etiquetación manual: agentes humanos etiquetan manualmente cada término según un conjunto de reglas consensuadas entre los investigadores.

El modelo más efectivo, siempre que sea posible, es el automático. Pero presenta varios problemas:

- Se necesita un algoritmo fiable, que en general debe estar basado en un conjunto de datos previos de tamaño significativo.
- De lo anterior se deduce que cualquier algoritmo, sin los datos necesarios, sea incapaz de responder ante situaciones inéditas. Por ejemplo, la aparición de términos nuevos no previstos.
- Como en cualquier proceso automático, es necesario contar con un número de errores posibles que se darán por análisis válidos, con lo cual la revisión final manual es conveniente y, muchas veces, necesaria.
- La discriminación entre dos análisis diferentes para un mismo término puede ser indistinguible para un sistema automático. Por ejemplo, los casos de *como* (conjunción), *como* (adverbio) y *como* (primera persona del presente de indicativo del verbo *comer*) suelen ser indistinguibles para algunos analizadores automáticos.

Puesto que el objetivo de gran parte de los estudios basados en corpus lingüísticos suele ser, entre otros, el análisis de textos para posteriormente crear sistemas automatizados, consideramos adecuado recurrir, en esta primera fase, a la etiquetación manual. Sin embargo, el etiquetado manual resulta una labor pesada y propensa a errores. Por tanto, para el presente trabajo optamos por la creación de una herramienta que funciona como un asistente a la etiquetación manual, intentando facilitar de forma constante la creación de nuevos análisis y su revisión.

Cuando se realiza la marcación de corpus lingüísticos, sobre todo de naturaleza léxica, es usual que se repitan los mismos análisis en diferentes contextos. Por tanto, el sistema se ha diseñado para aprovechar esta característica y facilitar el trabajo.

#### 4.3.2. Modelo de aplicación

La aplicación permite el etiquetado de un conjunto de términos que se agruparán en unidades funcionales llamadas *documentos*. La estructura de nuestro estudio basado en los archivos de disponibilidad léxica ha determinado que consideremos como documentos cada una de las listas de palabras organizada por sujetos y centros de interés. Así, el documento 001 01 contiene las palabras que el sujeto 001 escribe en el centro de interés 01, el documento 001 02 las palabras escritas por el mismo sujeto en el centro de interés 02, el documento 004 10 aquellas palabras que el sujeto 004 escribe en el centro de interés 10 y, así, sucesivamente. La Tabla 3 muestra un ejemplo de un documento de trabajo.

Tabla 3. Documento de trabajo en *Annotate*. Sujeto 001, centro de interés 01

<b>Orden</b>	<b>Lema</b>
1	mano
2	pie
3	brazo
4	cerebro
5	pulmón
6	nariz
7	extremidad
8	ojo
9	boca
10	diente
11	pelo
12	oreja
13	culo
14	vagina

Un conjunto de documentos que comparten una estructura de análisis común constituye un *proyecto*. En nuestro caso, al conjunto de documentos compuesto por los textos que todos los sujetos actualizaron en cada uno de los centros de interés lo llamamos

*Proyecto CONSOLEX*. Los análisis han de ser configurables para cada proyecto y deben permitir la reutilización de los campos definidos para los análisis de otros proyectos para organizar y potenciar su coherencia.

*Annotate* permite la etiquetación de cada término pero, puesto que es habitual que los términos se repitan, se utiliza el concepto de ‘modelo’, gracias al cual se pueden consultar los análisis previos y reutilizarlos en el análisis actual. De este modo, cuando se analiza por primera vez un término, este queda almacenado en la base de datos como *modelo1*, de manera que cuando vuelva a aparecer el mismo término en otro documento, el usuario puede seleccionar el modelo almacenado previamente sin necesidad de realizar otra vez el análisis. Del mismo modo, si fuera necesario introducir alguna modificación en el modelo existente (por ejemplo, añadir una marca de error ortográfico o similar), el sistema almacenaría el nuevo análisis como *modelo2*, de forma que, ante una nueva aparición del mismo término, el usuario podría elegir entre el *modelo1* o el *modelo2* y, así, sucesivamente.

Estos conceptos modelan *Annotate* y lo convierten en un asistente a la etiquetación manual altamente configurable.

#### 4.3.3. Características

*Annotate* se construye sobre una plataforma Web para facilitar su utilización por múltiples usuarios y el acceso global. La aplicación se aloja en un servidor ubicado en el *Servicio Central de Informática de la Universidad de Málaga*. Actualmente, y por motivos de seguridad y control, es preciso contar con autorización para usar la aplicación. El acceso a la herramienta se realiza a través del enlace <http://linguae.filosofia.uma.es/Tools/Annotate> [última consulta, 21 de abril de 2020] y el

usuario interesado en crear nuevos proyectos puede obtener acceso por medio del contacto [amavila@uma.es](mailto:amavila@uma.es).

Toda la implementación se ha realizado con el objetivo de la mayor eficiencia y la menor transferencia de datos posible, lo que posibilita el uso de la aplicación sobre canales con poca capacidad de transferencia y una alta movilidad.

El sistema se estructura en dos niveles principales:

- Descripción de campos: donde se describen, configuran y ordenan los tipos de datos que podrán ser recogidos en los diferentes análisis. A cada variable se le otorga un nombre y se le acompaña de una descripción aclaratoria sobre su naturaleza. El tipo de datos que contiene cada variable puede ser fijo o de opciones. En este último caso, las distintas opciones o variantes que puede elegir el usuario son consideradas los valores del campo.
- Proyectos: en los que se agrupan los datos que serán analizados con un conjunto de campos seleccionado. Dentro de los proyectos se encuentran los documentos que agrupan conjuntos de términos relacionados. En nuestro caso, el *Proyecto CONSOLEX* contiene las listas que los sujetos actualizaron en cada uno de los centros de interés propuestos. Para este Proyecto configuramos los campos que se muestran en la Tabla 4.

Tabla 4. Estructura de campos usada en el *Proyecto CONSOLEX*

<b>Nombre de la variable</b>	<b>Descripción</b>	<b>Valores</b>
Lema	<i>Lema</i>	
Realización	<i>Realización</i>	
Lema-Realización	<i>Coincidencia o variación entre Lema y Realización</i>	1. Coincidencia 2. No coincidencia
Categoría	<i>Categoría gramatical</i>	1. Nombre 2. Verbo 3. Adjetivo 4. Adverbio 5. Locución preposicional 6. Otros

		9. No aplica
Género	<i>Género gramatical</i>	1. Masculino 2. Femenino 3. Masculino y femenino 4. Común 5. No procede 9. No aplica
Número	<i>Número gramatical</i>	1. Singular 2. Plural 3. No procede 9. No aplica
Tecnicismo	<i>Naturaleza especializada del término</i>	1. No 2. Semiespecializado 3. Semiespecializado NO DLE 4. Semiespecializado NO DLE Nombre propio 5. Semiespecializado NO DLE Siglas 6. Especializado 7. No procede 9. No aplica
Arcaísmo	<i>Término arcaico u obsoleto</i>	1. No 2. Poco usado 3. Desusado 4. Anticuado 5. No procede 9. No aplica
Dialectalismo	<i>Término asociado a zona geográfica concreta</i>	1. No 2. Urbano malagueño 3. Provincial malagueño 4. Provincial no malagueño 5. Dialectal andaluz 6. Otras procedencias 7. No procede 9. No aplica
Error ortográfico	<i>Tipo de error ortográfico</i>	1. No 2. Acentuación 3. Grafías c/z+e,i 4. G/J 5. S/X/C/K/Q 6. H 7. B/V 8. LL/Y/Ñ 9. No aplica 10. Grafía original extranjera 11. + de un error 12. Otros 13. No procede
Formación	<i>Formación o estructura morfológica</i>	1. Simple 2. derivada sufijal 3. Derivada prefijal 4. Derivada prefijal y sufijal

		5. Composición 6. Composición culta 7. Truncación 8. Antonomasia 9. No aplica 10. Compuestos reales 11. Otros 12. No procede
Intencionalidad	<i>Valoración/intencionalidad</i>	1. No 2. Eufemismo 3. Malsonante 4. Irónico 5. Figurado 6. Peyorativo 7. Festivo 8. Despectivo 9. No aplica 10. No procede
Préstamo	<i>Préstamo de otras lenguas</i>	1. No 2. Extranjerismo aceptado y adaptado DLE 3. Extranjerismo aceptado DLE cursiva 4. Extranjerismo 5. Extranjerismo Nombre Propio 6. Extranjerismo Siglas 7. No procede 9. No aplica
Uso	<i>Nivel de uso y registro</i>	1. No 2. Infantil 3. Jergal 4. Culto 5. Vulgar 6. Rural 7. Popular 8. Coloquial 9. No aplica 10. No procede
Centralidad	<i>Grado de centralidad con el centro de interés propuesto</i>	1. Nuclear 2. Derivaciones 3. Asociaciones colectivas 4. Asociaciones individuales

Cada documento puede ser anotado con información común al conjunto de términos que aglutina y permite organizar la etiquetación de datos al nivel más conveniente.

Para el estudio que aquí presentamos, distinguimos diferentes grados de especialización para cada vocablo (Guerrero Ramos y Pérez Lagos, 2018): 1. *No*: el

elemento analizado pertenece al léxico común; 2. *Semiespecializado*: si en el DLE aparece acompañado de marca técnica; 3. *Semiespecializado NO DLE*: si pertenece en su origen a algún campo de especialidad, pero no está recogido en el DLE. 4. *Semiespecializado NO DLE NombrePropio*: Nombre propio que pertenece en su origen a algún campo de especialidad, pero no está recogido en el DLE. 5. *Semiespecializado NO DLE Siglas*: sigla que pertenece en su origen a algún campo de especialidad, pero no está recogido en el DLE. 6. *Especializado*: no aparece en los diccionarios generales y solo es explicada en diccionarios técnico-científicos.

#### 4.3.4. Posibilidades

El objetivo de *Annotate* es conseguir el etiquetado de términos mediante una estructura de etiquetación específica para cada proyecto. Como hemos señalado, su organización se basa en una estructura predefinida (proyectos-documentos) que permite etiquetados a diferentes niveles: proyecto, documento y/o término.

Tal y como se muestra en la Tabla 4, el conjunto de etiquetas que se emplea es configurable y puede cubrir cualquier ámbito: información etimológica, contextual de la recogida de datos, gramatical específica de cada término, etc.

Los proyectos que compartan conjuntos de etiquetas pueden generar información compatible, restringiendo los datos extraídos de la aplicación a cuantos les sean comunes. De esta manera se puede comparar, integrar o correlacionar conjuntos de datos provenientes de diferentes proyectos, con lo que esto supone de enriquecimiento y aprendizaje continuo de la herramienta *Annotate*.

#### 4.3.5. Importación y exportación de datos

El sistema permite la construcción manual de proyectos, documentos y líneas de análisis, pero esta forma de construir la información suele ser poco productiva y propensa a errores. Puesto que los datos suelen estar preparados para el uso con herramientas previamente diseñadas —como LexiDisp en el caso de los estudios de disponibilidad léxica—, se proporciona un mecanismo de importación de datos compatible con otros sistemas que crea automáticamente la estructura de documentos y las líneas de análisis en un proyecto. Así, la única condición aplicable a los datos de entrada es que tienen que estar almacenados en formato TXT y su estructura ha de ser la que exige LexiDisp (<http://www.linguas.net/LinkClick.aspx?fileticket=f2GkSGZ34ts%3d&tabid=73&mid=429&language=es-ES> [última consulta, 21 de abril de 2020]). A continuación, se muestra un ejemplo de la estructura y organización de los ficheros de entrada a *Annotate*.

*001 01 mano, pie, brazo, cerebro, pulmón, nariz, extremidad, ojo, boca, diente, pelo, oreja, culo, vagina*

Como puede observarse, la única diferencia con los ficheros de entrada a LexiDisp es que no son necesarios los cinco dígitos iniciales que representaban las variables socioculturales que manejaba aquel programa. Nuestros ficheros están asociados a una base de datos sociológica independiente que contiene información sobre cada informante. El resto de la organización interna de los ficheros es idéntica: *001* hace referencia al número de identificación del informante, *01* se refiere al centro de interés y, a continuación y separadas por comas, aparecen las formas lematizadas actualizadas por cada sujeto en sus listas. Una vez importados los ficheros de datos, estos quedan preparados para el análisis.

La exportación de datos se realiza integrando la información proporcionada por proyectos, documentos y líneas de análisis en un archivo CSV, compatible con otras herramientas de análisis exteriores como SPSS, Access, Excel, entre otras.

## **5. Análisis**

Las hipótesis de este estudio están basadas en que la posición de los hablantes y sus características personales (edad, género, etc.) son factores susceptibles de predecir: 1) el número de palabras actualizadas por cada centro o Índice de Amplitud Léxica (IAL); 2) la naturaleza relativa del vocabulario actualizado medida en el promedio de palabras usadas por un informante y que no comparte con los demás para un centro de interés dado o Índice de Descentralización Léxica (IDL) y 3) el número de vocablos especializados que contienen los listados individuales (IVE).

Cuanto más elevado es el número de palabras actualizadas por un hablante, mayor es la probabilidad de que exista un subconjunto cuya naturaleza no coincida con las seleccionadas por los demás. Esto también tiene que ver con la disponibilidad de cada palabra en cada individuo, pues las palabras más accesibles son las que se actualizan antes y resultan las más frecuentes para el colectivo. Las dos primeras hipótesis están relacionadas: cuanto más elevado es el número de palabras actualizadas por un hablante, mayor es la probabilidad de que exista un subconjunto de palabras que no coincida con las palabras seleccionadas por el resto. Además, esto también tiene que ver con la disponibilidad de cada palabra en cada individuo, puesto que las palabras más accesibles son las que se actualizan antes y resultan las más frecuentes desde el punto de vista colectivo. La correlación entre ambas variables es positiva, fuerte y significativa ( $R = 0.988$ , sig. = 0.000,  $N = 36$ ). De hecho, se trata del reflejo estadístico de una constante

cualitativa conocida en el campo de la lexicología: los marcos cognitivos están compuestos por un núcleo central de elementos léxicos que comparten muchas propiedades con el prototipo de la clase en cuestión (*cabeza* o *pierna* en el vocabulario del cuerpo humano y *gorrión* o *canario* en el de las aves, por ejemplo). Conforme nos alejamos de ese centro de elementos muy característicos y muy frecuentes, las palabras pierden frecuencia de uso (y ganan en especificidad), al ser prototípicas en menor grado (como *esteatopigia* o *lipodistrofia* en el vocabulario relacionado con el cuerpo humano y *pingüino* o *alcatraz* con el de las aves).

Para comprobar las hipótesis se han usado tres variables léxicas que miden los datos obtenidos en la investigación en función de: 1) el número de palabras por hablante y centro de interés (Índice de Amplitud Léxica; IAL). Para este trabajo se consideró el número total de palabras en los veinte centros de interés; 2) la descentralización de las palabras actualizadas por cada hablante en cada centro de interés (IDL). El IDL es una función del vocabulario del sujeto en un centro de interés y de su grado de compatibilidad con el vocabulario de los demás sujetos en ese centro. Para este estudio se ha empleado la suma de los índices en cada hablante y en los veinte centros; 3) el número de vocablos especializados contenidos en las listas (IVE); se analizaron con *Annotate* todas las palabras contenidas en los veinte centros de interés y se obtuvieron 726 términos marcados con algún grado de especialización, lo que supone casi un 3% del total de formas (24 634).

Las variables de hablante introducidas fueron, por una parte, la ocupación (0-3) y los estudios (en años cursados), porque son indicadores tradicionales de la clase social; a ellos se añadieron, por otra parte, los años de estudios del padre, puesto que, como se sabe, la posición y, en particular, la educación de los padres, influyen en la de los hijos, ya que condicionan el contexto educacional. La edad se considera importante puesto que revela la existencia de cambios lingüísticos en curso y estratificaciones generacionales

estables. La ocupación se considera asimismo como un indicador de la posición social, puesto que las profesiones se asocian con distintos grados de prestigio en la comunidad. El índice de clase social, por último, trata conjuntamente los efectos de la educación, la ocupación y la renta familiar (03).

### 5.1. Correlación

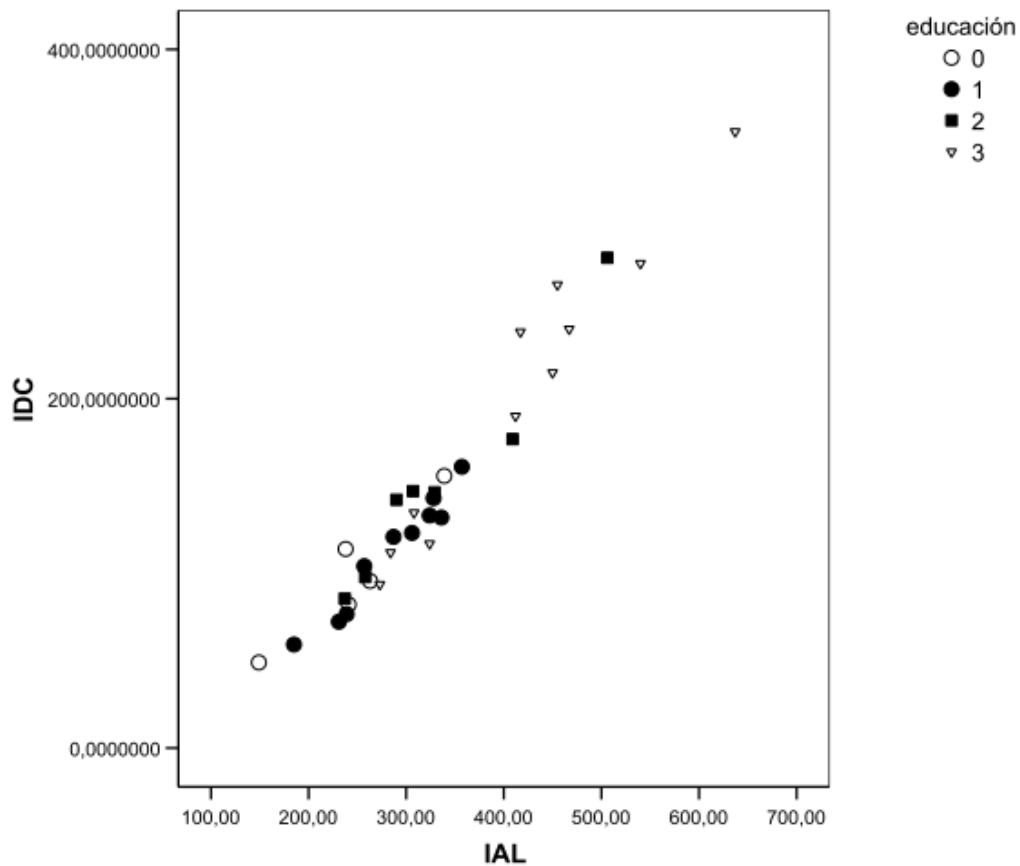
La correlación de las variables léxicas con las variables de hablante consideradas se muestra en la Tabla 5. La educación del hablante (en años cursados), por sí misma, se relaciona fuerte y linealmente con las variables léxicas. Otras variables a tener en cuenta son: clase social, grado de contacto del individuo con la información (*Einform*), ocupación, años de estudio del padre y, en menor medida, edad del sujeto. *Einform* es una variable compleja creada mediante la suma de los valores obtenidos por cada informante en tres escalas que miden su relación con algunos de sus hábitos de consumo cultural: lectura de prensa nacional (escala de 0 a 4, donde 0 = nunca y 4 = todos los días), lectura de libros al año (0 = ninguno hasta 4 = más de 20) y consumo de programas televisivos del corazón y similares (0 = todos los días hasta 4 = nunca). *Einform* produce para cada individuo un valor situado entre 0 y 12 que indica el grado y características de su exposición a la información. En nuestro estudio, esta variable no está correlacionada con la educación de los informantes ( $R = 0.261$ ,  $\text{sig.} = 0.732$ ,  $N = 36$ ): hay sujetos con escasa educación formal que obtienen puntuaciones muy altas en esta escala y, a la inversa, la educación superior no garantiza alta puntuación en esta escala compleja.

Tabla 5. Correlaciones

Rho de Spearman		Ocupación	Clase social	Estudios del padre	Einform	Educación	Edad
Índice de amplitud léxica (IAL)	Coefficiente de correlación	,469**	,497**	,480**	0,243	,575**	-0,102
	Sig. (bilateral)	0,008	0,003	0,005	0,180	0,001	0,556
	N	36	36	36	36	36	36
Índice de descentralización (IDL)	Coefficiente de correlación	,436*	,505**	,510**	0,258	,543**	-0,110
	Sig. (bilateral)	0,011	0,003	0,002	0,166	0,001	0,521
	N	36	36	36	36	36	36
Índice de vocablos especializados (IVE)	Coefficiente de correlación	,484*	,503**	,594*	,719**	,854**	,554*
	Sig. (bilateral)	0,011	0,003	0,013	0,001	0,000	0,017
	N	36	36	36	36	36	36

El Gráfico 1 representa la posición de los hablantes en el espacio definido por las dos variables léxicas con mayor relación: la amplitud léxica (IAL) y la descentralización (IDL). Los hablantes aparecen marcados como incluidos en un grupo educacional recodificado (03). La relación entre la educación y las dos variables léxicas es positiva.

Gráfico 1. Situación de los hablantes del grupo estudiado en Málaga en el espacio de interacción de las variables léxicas



## 5.2 Patrones sociolingüísticos

Se recodificaron los años de estudios de los hablantes y se creó una variable (*educación*) que distingue cuatro grupos educacionales (03). La idea era comparar los promedios de uso de las variables léxicas en cada uno de los grupos, para comprobar que, tal y como sugiere el análisis anterior, el grupo analizado se estratifica educacionalmente.

Las diferencias de medias en el uso de IAL e IVE son significativas entre los grupos educacionales (Tabla 6).

Tabla 6. Diferencias educacionales en IAL e IVE

	IAL			IVE		
	Media	Desy.	N	Media	Desy.	N
0	24600	6789	6	35	82	6
1	28500	5519	11	123	133	11
2	33371	9406	8	235	143	8
3	41518	11291	11	321	233	11
Total	33281	10630	36	726	210	36

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
IAL	Inter-grupos	135179,844	3	45059,948	5,771	0,003
	Intra-grupos	226441,065	29	7808,313		
	Total	361620,909	32			
IVE	Inter-grupos	1235,765	3	6543,65	2,765	0,000
	Intra-grupos	43212,654	14	2343,76		
	Total	98763,654	17			

Recorte rectangular

Existen, pues, diferencias significativas entre las medias de los diferentes grupos educacionales (03) en el uso del IAL e IVE. Sin embargo, conviene saber cuáles de estas diferencias son significativas y cuáles no. Los contrastes entre las medias de los grupos (*test de Tukey*) revelaron que las diferencias significativas se establecen entre los hablantes que han cursado bachillerato y nivel universitario (2 y 3) y el resto (0 y 1). Una vez recodificados los dos grupos, el análisis de las diferencias de medias para dos muestras independientes (*ttest*) muestra que estas son significativas en las dos variables léxicas consideradas aquí (Tabla 7).

Tabla 7. Diferencias entre los grupos educacionales mediobajo y medioalto

	Grupos Educcionales	N	Media	Desy.	Error Típ.
IAL	0 + 1	17	2722300	6030991	1557195
	2 + 3	19	3833240	11087367	2613317
IVE	0 + 1	17	334	12039	3807
	2 + 3	19	562	14205	5022

ttest: IAL:  $t(27,071) = -3.665$ , sig. = 0.001; IVE:  $t(13,810) = -5.303$ , sig. = 0.000

## 6. Discusión

1. La posición social de partida parece condicionar las oportunidades de ocupar situaciones ventajosas en los mercados socioculturales y lingüísticos. Este escenario está bastante cerca del reflejado en la teoría de los códigos (restringido y elaborado) de Basil Bernstein (1971 y 1975) y supone, por un lado, el acceso de los hablantes de los estratos más bajos a estructuras gramaticales, estrategias discursivas y léxico restringidos (esto es, en nuestro caso, menor amplitud léxica, léxico muy centralizado y, en consecuencia, poco especializado); por otro lado, los hablantes de los estratos más altos disponen de la capacidad de utilizar además de los recursos del código restringido, los que integran el código elaborado (es decir, aquí, el vocabulario más amplio, más descentralizado y con mayor grado de especificidad). El código elaborado ocupa un papel central en la estructura social. La lengua ejemplar, el lenguaje científico y el literario, el administrativo y el político, forman parte del código elaborado en el que también se desenvuelve el sistema escolar. El aparato educativo suele fomentar solo el código elaborado y, por ello, los alumnos procedentes de las clases bajas están en inferioridad de condiciones académicas de partida. Este hecho es causa de un alto índice de fracaso escolar y motivo de que el déficit lingüístico comporte otro cognoscitivo. La escuela y, más tarde, la universidad, tienden a reproducir esta situación de modo que es difícil evitar la imagen clásica del *círculo infernal de la pobreza*. La representación ideológica subyacente es que las variedades no estándar representan un obstáculo para el progreso social mismo, puesto que afectan a la comunicación, especialmente en la escuela (Villena, 2012: 165). Como es de suponer, la propuesta de Bernstein abrió un campo de aplicación sociolingüística que no fue desaprovechado en absoluto.

2. Hace años que el *paradigma cuantitativo* del variacionismo sociolingüístico (Weydt y Schlieben-Lange, 1981) propuso una postura más *optimista* proclive a la interpretación, entendida como complemento —o directamente como alternativa— al corsé estricto al que obligaban perspectivas *pesimistas* de las variedades vernaculares como la propuesta por Bernstein. Este giro interpretativo fue el resultado acumulado de varias corrientes que, consideradas en su momento como anecdóticas, demostraron con el paso del tiempo tener sentido y permitieron reconstruir una corriente variacionista anticonformista de fundamento crítico y emancipador cuyo origen y trayectoria no se identifica con la línea dominante basada, fundamentalmente, en una ideología de terapéutica social y lingüística (Bourdieu 1984; Sankoff, 1988; Williams 1992; Villena, 1992). Esta ideología de carácter positivo y abierto está en el origen del variacionismo y justifica la ampliación y reforma de los objetivos y métodos de la lingüística llamada *asocial* (terapéutica lingüística). Por supuesto, aspira a la reconstrucción de las condiciones sociales de igualdad en ámbitos claves como la educación (terapéutica social). No olvidemos que la finalidad de la sociolingüística en sus orígenes fue la liberación de los individuos del control y la opresión, en la línea del pensamiento, por ejemplo, de Habermas. Los síntomas principales de esta nueva situación fueron detectados desde hace tiempo (Romaine, 1982; Milroy, 1992; Auer y Hinskens, 1996): los individuos parecen demostrar cierta capacidad para vencer una estructura social y comunicativa polarizada y fuertemente determinada por las condiciones socioeconómicas desiguales.

Los resultados obtenidos en nuestra investigación tienden a reforzar la línea, en cierto modo también optimista de los planteamientos fundadores de la sociolingüística (Labov, Shuy, etc.) que surgieron como una respuesta crítica ante los problemas de comunicación: la exposición, a lo largo de la biografía personal de los hablantes, a las

fuentes de información y de conocimiento especializado y técnico aumenta la ampliación de vocabulario y, en general, la capacidad léxica.

## **7. Conclusiones**

A lo largo de este trabajo hemos analizado el efecto conjunto de la estratificación socioeconómica y de las diferencias de atributos personales de los hablantes en la disponibilidad léxica en una muestra representativa con afijación proporcional de la población de la ciudad de Málaga. El estudio ha consistido en 1) la realización de pruebas específicas de disponibilidad por centros de interés; 2) el estudio de cada término obtenido gracias al empleo del asistente semiautomático de lematización *Annotate* y, 3) el análisis cuantitativo de los datos. Todo ello nos ha servido para observar que existe un condicionamiento estratificacional del vocabulario a disposición de los hablantes, que los diferencia en el número, la adecuación y la naturaleza de las palabras que pueden evocar en ciertos contextos conceptuales y semántico-situacionales.

Gracias a la interpretación del análisis realizado, nos hemos situado en los orígenes de la disciplina sociolingüística para entender la estratificación de la disponibilidad léxica en la muestra analizada. Se ha demostrado que la posición de los hablantes y sus características personales predicen, en efecto, cierto grado de variación léxica. En concreto, las variables con más capacidad de predicción de las variables dependientes usadas han sido la educación, la clase social, los estudios del padre, la ocupación, la edad de los informantes y la escala de consumo cultural. Si bien es cierto que las cinco primeras variables independientes corroboran las teorías deterministas de B. Bernstein —y nos sitúan en el nivel de la estructura social, teoría de la estratificación

estructural, semejanza estructural— la posibilidad de que el consumo cultural personal se presente como una variable predictora —nivel individual, historia social, semejanza intencionada— es, hasta cierto punto, original y esperanzadora.

El Índice de Vocablos Especializados (IVE) es una variable fuertemente relacionada con la educación de los individuos; sin embargo, no resulta menos intensa la relación con la escala de acceso al consumo cultural. La exposición, a lo largo de la biografía personal de los hablantes, a las fuentes de información y de conocimiento especializado y técnico aumenta la ampliación de vocabulario y, en general, la capacidad léxica. En consecuencia, la desigualdad educacional no determina fatalmente el acceso a las fuentes de conocimiento y de ampliación del vocabulario. La obtención sistemática de vías de información tras el período formativo escolar no es patrimonio exclusivo de los hablantes de mayor capacidad léxica en su origen. Esta, como otras capacidades, se puede deteriorar por falta de cultivo o puede adquirirse y mejorarse gracias a la dedicación y a la actitud personal positiva.

## **Bibliografía**

- Auer, Peter y Frans Hinskens. 1996. "The convergence and divergence of dialects in Europe. New and not so new developments in an old area". En *Konvergenz und Divergenz von Dialekten in Europa*. (Sociolingüística 10), ed. por Peter Auer, Frans Hinskens y Klaus J. Mattheier, 1-30. Tübingen: Newmwyer.
- Ávila, Antonio M. 2006. *Léxico disponible de los estudiantes preuniversitarios malagueños*. Málaga: SPICUM.
- Ávila, Antonio M. 2016. "El léxico disponible y la enseñanza del español. Propuesta de selección léxica basada en la teoría de los conjuntos difusos." *Journal of Spanish Language Teaching*, DOI: 10.1080/23247797.2016.1163038.
- Ávila, Antonio M. 2017. "Entre limonetas y malaguitas: Estudio del léxico divergente de proximidad. Nuevos datos del Proyecto CONVERLEX-Málaga." *Hispania* 100/4: 538-553.
- Ávila, Antonio M. 2019. "La integración sociocultural de los grupos vulnerables de migrantes desde el aula de lenguas. Propuesta de actuación desde la dialectología social en la ciudad de Málaga." *Lengua y migración / Language and Migration* 11:1: 73-97.
- Ávila Antonio M. y José M. Sánchez. 2010. "La posición de los vocablos en el cálculo del índice de disponibilidad léxica: procesos de reentrada en las listas del léxico disponible de la ciudad de Málaga." *ELUA* 15: 35-82.

- Ávila, Antonio M. y, Juan A. Villena. 2010. *Variación social del léxico disponible en la ciudad de Málaga. Diccionario y análisis*. Málaga: Sarriá.
- Bernstein, Basil. 1971. *Class, Codes and Control: Theoretical Studies Towards a Sociology of Language*. Londres: Routledge y Kegan Paul.
- Bernstein, Basil. 1975. *Class, Codes and Control: Towards a Theory of Educational Transmissions*. Londres: Routledge y Kegan Paul.
- Bourdieu, Pierre. 1984. "Capital et marché linguistiques." *Linguistische Berichte* 90: 3-24.
- Childs, Becky y Christine Mallinson. 2006. "The significance of lexical items in the construction of ethnolinguistic identity." *American Speech* 81: 3-30.
- Geeraerts, Dirk. 2010. *Theories of Lexical Semantics*. Oxford: Oxford Linguistics.
- Geeraerts, Dirk, Stefan Grondelaers y Peter Bakema. 1994. *The Structure of Lexical Variation. Meaning, Naming, and Context*. Berlín: De Gruyter.
- Guerrero, Gloria y Manuel F. Pérez. 2018. "Vocabulario especializado en el léxico disponible de los hablantes malagueños." *Tonos Digital*, 34.
- Labov, William y Wendell A. Harris. 1986. "De facto segregation of black and white vernaculars" En *Diversity and Diachrony*, ed. por David Sankoff, 1-25. Ámsterdam-Filadelfia: John Benjamins.
- Milroy, James. 1992. *Linguistic Variation and Change: On the Historical Sociolinguistics of English*. Oxford: Blackwell.
- Milroy, Lesley y James Milroy. 1992. "Social network and social class: Toward an integrated sociolinguistic model." *Language in Society* 21(1): 1-26.
- Romaine, Suzanne. 1982. "On the epistemological status of sociolinguistic theory." En *Socio-historical Linguistics Its Nature and Status*, Cambridge: Cambridge University Press, 239-289.
- Sankoff, David. 1988. "Sociolinguistic and Syntactic Variation." En *Linguistics: The Cambridge Survey*, vol. 4: *Language: The Sociocultural Context*, ed. por Frederick J. Newmeyer, 140-161. Nueva York: Cambridge University Press.
- Sánchez-Sáez, Marta. 2019. *Centros de interés y capacidad asociativa de las palabras*. Sevilla: Universidad.
- Trigo, Ester, Manuel F. Romero e Inmaculada C. Santos. 2019. "Aproximación empírica desde la disponibilidad léxica a la influencia de los factores sociolingüísticos en el dominio ortográfico." *Cultura y educación* 31(4): 1-31.
- Villena, Juan A. 1992. *Fundamentos del pensamiento social sobre el lenguaje*. Málaga: Ágora.
- Villena, Juan A. 2012. "La investigación sociolingüística de la comunidad de habla: el origen inconformista de la dialectología social." *Revista de Filología* 30: 155-176.
- Villena, Juan A. y Antonio M. Ávila. 2014. "Dialect stability and divergence in southern Spain." En *Stability and Divergence in Language Contact: Factors and Mechanisms*, ed. por Kurt Braunmüller, Steffen Höder y Karoline Köhl, 207-238. Ámsterdam: John Benjamins.
- Villena, Juan A., Antonio M. Ávila y María C. von Essen. 2017. "Efecto de la estratificación, la red social y las variables de pequeña escala en la variación léxica. Proyecto de investigación sobre la convergencia del léxico dialectal en la ciudad de Málaga (Converlex)". En *Léxico Español Actual V*, ed. por Luis Luque y Rocío Luque, 209-233. Università Ca' Foscari Venezia: Venecia.
- Weydt, Harald y Brigitte Schlieben-Lange. 1981. "Wie realistisch sind Variationsgrammatiken?" En *Logos Semantikos. Studia Linguistica in Honorem E. Coseriu*, vol. 5, ed. por Wolf Dietrich y Horst Geckeler, 117-145. Madrid, Berlín y Nueva York: Gredos/de Gruyter.
- Williams, Glyn. 1992. *Sociolinguistics: A Sociological Critique*. Londres y Nueva York: Routledge.