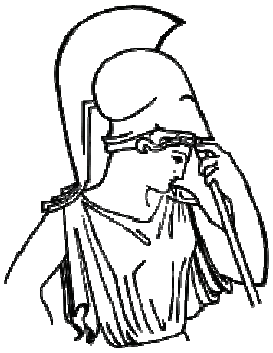
# UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

**FACULTAD DE FILOSOFÍA Y LETRAS**
**COLEGIO DE LETRAS HISPÁNICAS**

# LINGÜÍSTICA SOCIALMENTE REALISTA ENSAYO DE UN MÉTODO SOCIOLINGÜÍSTICO EN ESPAÑOL MEXICANO

## TESIS

QUE PARA OBTENER EL TÍTULO DE:
**LICENCIADA EN LENGUA Y LITERATURAS HISPÁNICAS**



PRESENTA:
**YAEL JERUSALEM RODRÍGUEZ ZAMORA**

ASESOR:
MTRO. JULIO CÉSAR SERRANO MORALES

MÉXICO, D.F. MAYO DE 2012

*A mi mamá*

*A mis hermanos*

*"¿Quién, si ella no,*

*pudo fraguar este universo insigne*

*que nace como un héroe en tu boca?"*

*José Gorostiza*

# ÍNDICE

# AGRADECIMIENTOS

He llegado a la parte de la que se habla mucho cuando apenas se va a empezar la tesis, los famosísimos agradecimientos, por fin.

Gracias le doy a toda mi familia: a mi mamá por darme apoyo, a mis tres hermosos hermanos (mis alegrías) que siempre estuvieron a mi lado, a mi abue que ya huía de mis lecturas "de intelectuales" como ella les decía, a mi papá, a Carlos, a mis tíos, tías, primos y primas. Saben que los quiero y tengan por seguro que a esta tesis sí se le entiende no como a la de Pepe.

Gracias también a mis amigos que dividiré en grandes secciones, no porque sean poco importantes sino porque soy muy torpe y no quisiera dejar a nadie fuera, estas secciones se mencionarán en orden cronológico, primeramente me quito el sombrero ante mis amigos del Colegio Madrid, ellos me han acompañado desde pequeña y han sabido cómo hacerme reír y ver la vida desde un punto chispeante; luego de que dejé el Madrid y entré a la Facultad, en mi carrera conocí a tres muñequitas que me hicieron amable mi transitar por los turbulentos momentos de juventud, y para terminar (porque fue a los que conocí al final) están mis amigos de otras carreras a los que les debo inolvidables momentos.

Además de agradecer a mi familia y amigos agradezco a mis maestros que me impulsaron en la vida y en este trabajo dentro de los que resaltaré a Julio mi asesor quien además de ser mi maestro es y será siempre mi amigo.

# INTRODUCCIÓN

El presente trabajo es una tesis de carácter sociolingüístico sobre el levantamiento de datos de la lengua en uso, esto es, de la lengua real. No hay antecedentes de algún trabajo que se asemeje a este en México, es por esto que se subtitula *Ensayo sobre un método sociolingüístico en español mexicano*, pues aún no se ha aplicado un método con estas características, que esté centrado en la realidad de la lengua. Este ensayo es una primera respuesta a la pregunta: ¿Cómo se podría levantar un corpus *realista* para el español de la ciudad de México, tanto leído como hablado?

En el primer capítulo se aborda y se aclara a grandes rasgos el concepto de sociolingüística variacionista, el interés que tiene esta por la lengua vernácula y la importancia del dato socialmente realista, aclarando la importancia que tiene el cambio en la sociedad, el cual comporta también un cambio lingüístico.

En el capítulo dos se explica a detalle la metodología que se llevó a cabo para la obtención de los datos realistas. Este método podría aplicarse con un número mayor de informantes de distintas edades, estratos sociales y regiones geográficas del país y, haciendo recolecciones cada determinado tiempo, poder contar con un registro diacrónico y sociolingüístico del español mexicano bastante fidedigno.

El tercer capítulo es un análisis del léxico a partir de los datos que se obtuvieron con la metodología propuesta. Este capítulo aborda porcentajes comparativos que revelan en cifras lo que se le dedica al habla, a la escritura y a la lectura una persona en un día común. Estos resultados revelan la realidad lingüística que puede ser de mucha utilidad para la creación de corpora realistas y concretos.

Finalmente, se presenta un capítulo de conclusiones generales, la bibliografía y dos apéndices: el primero sobre el sistema de transcripción empleado y el segundo con un ejemplo de uso del programa de análisis léxico *Ant Conc*.

# CAPÍTULO 1

# DISCUSIÓN TEÓRICO-METODOLÓGICA: SOCIOLINGÜÍSTICA Y LA NATURALEZA DEL CORPUS SOCIOLINGUÍSTICO

## 1.1. ¿HOMOGENEIDAD DE LA LENGUA?

El concepto de *langue* para Saussure (1916 [1968]) constituye el lado social del lenguaje, sin embargo, la lengua ha sido tradicionalmente considerada como entidad homogénea. La lingüística de la primera mitad del siglo XX estudió la lengua como hechos lingüísticos sin que datos "de fuera" tomaran parte en el análisis; estos datos externos constituyen el contexto de la lengua, el ámbito donde se desarrolla; de esta manera el estudio de la lingüística quedaba reducido a la uniformidad que, se presupone, tiene ésta.

Las lenguas se organizan para comunicar a un grupo de personas integradas en una sociedad, su función principal es la comunicación por lo que se ubican en el plano de la actuación; ésta se desarrolla dentro de circunstancias variables, lo que conlleva a una actuación cambiante contenida en una lengua igualmente cambiante.

El conocimiento intersubjetivo de la lengua se halla precisamente en el habla de la vida cotidiana de los individuos --en el habla llamada también vernácula por no verse alterada por factores no naturales--, pero no podemos acceder a la lengua si no es a través del habla, esta es la paradoja saussureana.

Existen varios trabajos que sostienen que la heterogeneidad en la lengua es un asunto de multilingüismo y actuación[1], sin embargo la lengua es heterogénea por naturaleza, no se trata de un fenómeno homogéneo pues no se da con uniformidad; el concepto de idiolecto de Bloch y Hall[2] ocasiona tambaleos y caídas a la concepción saussureana establecida de *langue,* pues se pensaba que el entendimiento social era uniforme pero el surgimiento del concepto de idiolecto que, según Coseriu es: un sistema de la lengua como lo posee un individuo (cf. Bedmar 1995), comenzó a introducir muchas dudas al respecto. La sociolingüística variacionista, desde una perspectiva socialmente realista, pretende encontrar el orden que pueda haber en la variación y en los procesos de cambio lingüístico (cf. Weinreich, Labov & Herzog 1968).

Labov (1983) explica que un análisis de las relaciones estructurales de un sistema lingüístico no puede existir sin considerar también sus relaciones externas.

[1] Weinreich, Labov & Herzog (1968).

[2] "Conjunto de hábitos de un individuo en un tiempo dado".

Aunque existan lingüistas clave que suponen al contexto social como prescindible del análisis lingüístico (Bloomfield, Martinet, Troubetzkoy y Chomsky entre otros), encontramos también a muchos otros que poseen una perspectiva del lenguaje como un hecho social, como el lingüista Whitney, quien aclaraba: "el habla no es un bien personal sino un bien social; pertenece no al individuo, sino al miembro de la sociedad" (1901:404, citado en Labov 1983:326). En estos aspectos nos centramos en los siguientes apartados.

Algunas veces la lengua contiene periodos de cambio en donde ciertas variables toman impulso y crean cambios en el lenguaje; estos periodos de mayor inestabilidad son prueba de que la lengua es un material con propiedades de maleabilidad, en donde la rigidez de las reglas gramaticales se disuelve para dar cabida a los cambios en la lengua.

Los cambios fonéticos y léxicos muchas veces se dan para dar mayor identidad a un grupo, y se pueden producir desde distintos sectores de la sociedad –como lo muestran los cambios "desde arriba" y "desde abajo" documentados por Labov (1983)--lo que indica que el origen social de los cambios es irregular en la comunidad.

## 1.2. LA SOCIOLINGÜÍSTICA[3]

La sociolingüística es una disciplina asociada con otras como lo es la etnolingüística[4] o la sociología del lenguaje; disciplinas afines, pero que difieren en su campo de estudio.

La sociolingüística posee como objeto de estudio el lenguaje, su desarrollo, práctica y evolución en su entorno social; los objetivos que persigue, esto hay que aclararlo, son fundamentalmente lingüísticos, antes que sociales. Esto es, conlleva una visión focalizada en el lenguaje y cómo éste se ve afectado por los hechos y la estructura social. Los temas tratados desde esta disciplina son centrales para la lingüística general. Peter Trudgill en su artículo "Sociolinguística y sociolingüística" resalta que para Labov (1966) la sociolingüística es una *lingüística secular* (llamada también *lingüística verdadera* por Trudgill), un método de hacer lingüística, pues se centra en el lenguaje y en la naturaleza de los cambios lingüísticos, en su variabilidad y su estructura a partir de datos empíricos recogidos en el campo. Moreno Fernández (2010: 32) hace una clara distinción de disciplinas: "llamamos *sociolingüística* al estudio de la variación y el cambio

[3] La sociolingüística variacionista, también llamada *variacionismo,* nació en los años sesenta alrededor de la figura de W.Labov (1966;1972*b* y *c*). Tomado de Moreno Fernández (2009).

[4] También llamada etnografía del habla por Silva Corvalán ( 2001) y etnografía de la comunicación por Hymes (1974).

lingüístico en entornos urbanos y *sociología del lenguaje* al estudio de la presencia y el uso de la(s) lengua(s) en sociedades concretas". En el presente trabajo se considerará a la sociolingüística con las características expuestas.

A diferencia de la sociolingüística, la etnografía de la comunicación es una disciplina que se enfoca en el rol que desempeña la lengua en una comunidad, en una cultura que comprende la cosmovisión de un grupo social determinado; la etnolingüística se centra en el uso de la lengua para obtener conocimiento sobre la estructura social. Silva Corvalán (2001:11) muestra algunas de las preguntas comunes de esta disciplina: ¿Qué función tiene la conversación en la organización de la vida diaria? ¿Qué elementos (v.g., presupuestos culturales, conocimiento compartido) intervienen en la interpretación del significado de "lo que se dice"?

La sociología de la lengua (Fishman 1974) o sociología del lenguaje es otra disciplina que suele confundirse con la sociolingüística. Aunque compartan como objetivo el estudio de la actuación lingüística en situaciones reales (Silva-Corvalán 2001), la sociología de la lengua no tiene como objetivo el estudio de los fenómenos lingüísticos estrictos, no se centra en la lengua y en cómo se ve afectada ésta por el entorno social; su objetivo es el estudio de los fenómenos sociales que se ven afectados por la lengua.

## 1.3. EL ESTUDIO DE LA VARIACIÓN LINGÜÍSTICA

Los datos para el estudio de los cambios en la lengua deben constar con, entre muchas otras cosas, las pautas de prestigio y la autoevaluación de los hablantes. El estilo que toma cada individuo en su habla depende de la circunstancia en la que se encuentra, los factores circunstanciales dictan la elección del hablante por un código lingüístico u otro, la autoevaluación y pautas de prestigio del hablante proporcionan los datos para el estudio de la lengua en su contexto social. Además, W. Labov (1983: 262-263) se hace al respecto las siguientes preguntas:

> 1. ¿Cuál es la forma de las reglas lingüísticas? ¿Y qué constricciones deben imponérseles?
>
> 2. ¿Sobre qué formas subyacentes operan las reglas, y cómo podemos determinarlas con precisión en cada caso concreto?
>
> 3. ¿Cómo se combinan y ordenan las reglas en sistemas? ¿Y cómo se ordenan en el interior de dichos sistemas?
>
> 4. ¿Cómo se relacionan entre sí los sistemas en situaciones de bilingüismo y sistematicidad múltiple?
>
> 5. ¿Cómo cambian las reglas y los sistemas de reglas? ¿Cuál es el mecanismo de los procesos fundamentales de adquisición del lenguaje? ¿Cómo cambian las reglas a lo largo de la evolución lingüística general?

Dicho autor en su libro *Modelos sociolingüísticos* (1983), presenta como premisa para el estudio de la lengua a la variación que se lleva a cabo dentro de esta misma por el contexto social en el que se encuentra. En su capítulo ocho dice "El lenguaje es una forma de comportamiento social" y "(el lenguaje) es usado por los humanos en un contexto social para comunicarse sus necesidades, ideas y emociones unos a otros" (Labov 1983: 235). Un factor importante al estudiar la variación en el lenguaje es el tiempo, este factor es clave para el estudio del variacionismo pues, al hacer un análisis variacionista se puede proporcionar una explicación de los cambios lingüísticos que se forman en la lengua con el paso del tiempo. La lingüística variacionista pretende hacer un análisis pancrónico, es decir, tanto sincrónico como diacrónico, pues el variacionismo exige, a diferencia del estructuralismo, un análisis que conlleve la evolución que se da en la lengua a través del tiempo, que es lo que define el producto real.

La sociolingüística variacionista tiene como prioridad plasmar los hechos lingüísticos que se dan en la realidad, busca acercarse al lenguaje natural con el fin de obtener mayor información sobre la estructura de la lengua real:

> "Así pues, si queremos aprehender el *lenguaje* tenemos que examinar los datos del habla cotidiana lo más detallada y directamente posible, y caracterizar su relación con nuestras teorías gramaticales con la mayor

> precisión posible, corrigiendo y ajustando la teoría con el fin de que se adecúe (sic) al objeto de estudio." (Labov 1983:256).

Considera Labov que la palabra sociolingüística es tautológica, ya que no existe lenguaje sin sociedad. Su estudio debe tomar en cuenta tres aspectos que se presenta como preguntas (Labov 1983: 235): ¿Las funciones directiva y expresiva son determinantes importantes para el cambio? ¿Las reglas abstractas de la gramática pueden verse afectadas por las fuerzas sociales? ¿La revolución lingüística es completamente disfuncional o no? Todas estas interrogantes se han resuelto mediante el estudio de la lengua en función de la sociedad, de la que no es posible separarla.

A lo largo del tiempo las reglas de una sociedad cambian, las costumbres cambian y esto se ve reflejado en aspectos no verbales, en movimientos y gestos, aspectos que también forman parte de la situación comunicativa y que, junto con los aspectos verbales, se afectan mutuamente. Según Eugenio Coseriu (1958), un hecho lingüístico es el fruto tanto de una estructura social como de una estructura lingüística, para él es imposible concebir estas entidades por separado.

Los hablantes forjan su lenguaje dependiendo de lo que han oído que dictamina la sociedad ya como lengua de alto prestigio ya como lengua de poco estatus; para crear las reglas del habla ellos utilizan su intuición, una vez que han

forjado las reglas, los lingüistas las toman en cuenta como productos inconscientes e irreflexivos; así como el hablante utiliza la intuición para saber cómo es que se debe hablar dependiendo de la circunstancia, también selecciona palabras y el orden de éstas, es decir, puede utilizar una modalidad distinta del mismo lenguaje, con esta aseveración podemos decir que el cambio es evidente, pues hay propagación de variaciones en el lenguaje, ya que el lenguaje posee la característica de ser heterogéneo.

La sociolingüística variacionista, dice Gaetano Berruto (2010), es una lingüística que acepta y muestra diferentes realizaciones de la lengua en correlación con factores extralingüísticos, como los factores geográficos, sociales y de situación; por esto es que la **variable** lingüística representa la unidad mínima para el sociolingüista pues en ella se relacionan estrechamente la sociedad y el lenguaje; diría Berruto que es el punto que zurce al lenguaje y a la sociedad.

### 1.4. LA VARIABLE LINGÜÍSTICA COMO OBJETO DE ESTUDIO

La variación se muestra en maneras multiformes con significado adaptativo (Chambers 1995: 206-253).El mismo autor sostiene que la razón de la variación sociolingüística es establecer y mantener una identidad social.

Los cambios y las separaciones que se viven dentro de una lengua se dan para fortalecer una identidad de los que lo emplean, las variantes extralingüísticas crean variedades en el habla, éstas, a diferencia de lo que muchos piensan, tienen cierto orden en la heterogeneidad en la que existen. Berruto (2010) propone que para crear un modelo que delimite la variación lingüística hay que identificar las dimensiones de variación, estas dimensiones se clasifican en cuatro: la dimensión temporal (diacronía), la dimensión de la distribución geográfica de los que hablan (diatopía),la dimensión de la clase social de los que hablan (diastratia) y la dimensión de la situación comunicativa en la que se emplea (diafasia); a estas dimensiones también se les puede tratar como cuatro dimensiones: diacronía, diatopía, diastratia y diafasia. Todas éstas tienen una relación dinámica y se mueven a lo largo del tiempo, es decir, a lo largo del eje temporal de manera independiente, por esto es que dice que las dimensiones lingüísticas pueden representarse dentro de un cubo, sobreentendiendo la influencia del tiempo sobre todas las dimensiones. A continuación se presenta un diagrama que representa lo anterior por Berruto y posteriormente un diagrama mío adaptándolo en un cubo.

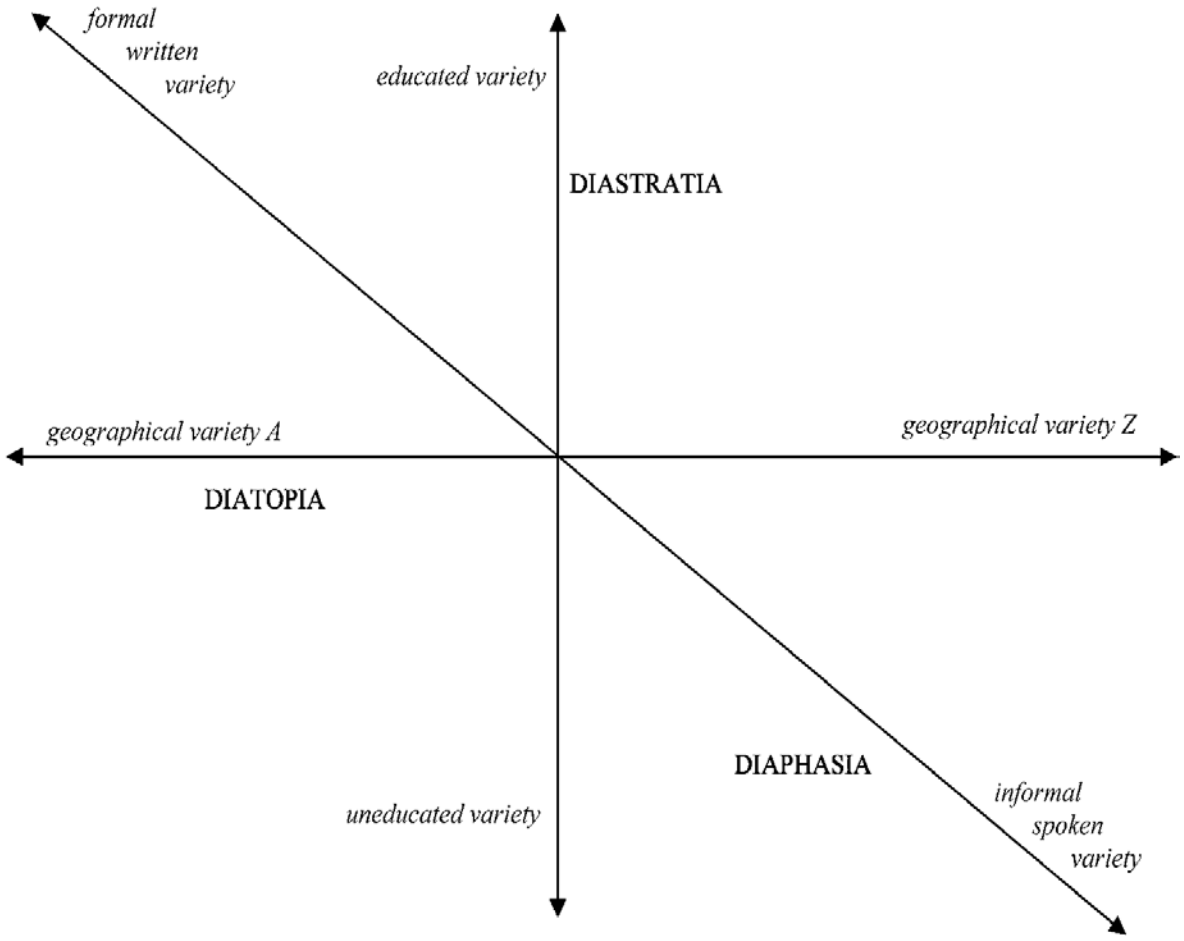



Fig. 1. *Arquitectura de una lengua como continuum multidimensional (tomado de Berruto 2010: 237).*

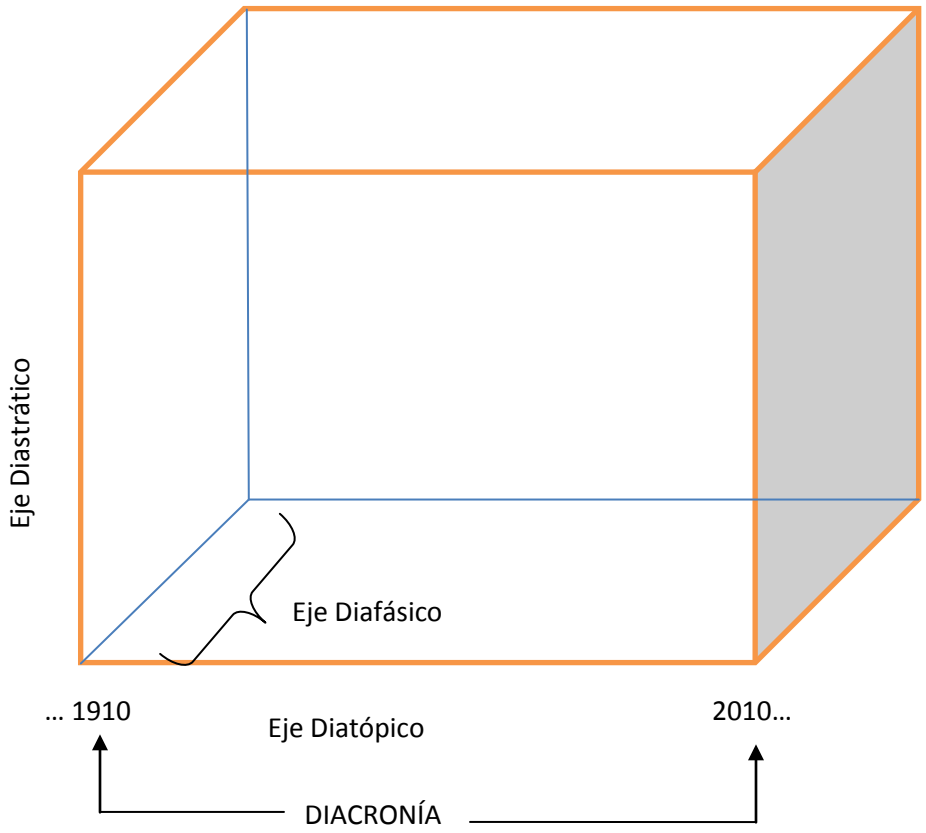



Fig. 2. *Cubo inspirado en Berruto (2010).*

A la sociolingüística actual (o "dialectología urbana") le interesa en particular la variación estilística o diafásica y la variación social (diastrática), mientras que la variación diatópica es tarea central de la geolingüística y la dialectología tradicional (Chambers & Trudgill 1994).Dialectología y sociolingüística tienen un interés compartido, que es el estudio de los cambios lingüísticos y cómo se difunden en el espacio y la sociedad.

## 1.5. RECOLECCIÓN DE DATOS SOCIOLINGÜÍSTICOS

Entre los desafíos que el sociolingüista debe enfrentar según Natalie Schilling-Estes (2007) están el cómo obtendrá los datos para el estudio; cómo encontrará una forma para que los hablantes se expresen de manera relajada y natural mientras toma el registro; cómo se harán grabaciones de buena calidad, especialmente en el exterior y no en un cuarto silencioso. Estos desafíos son aspectos derivados de la *paradoja del observador* que dice:

> El objetivo de la investigación lingüística de la comunidad ha de ser hallar cómo habla la gente cuando no está siendo sistemáticamente observada y sin embargo nosotros sólo podemos obtener tales datos mediante la observación sistemática (Moreno Fernández 1990: 29).

Explica además Schilling-Estes que un estudio sociolingüístico comienza con la selección de una población atendiendo a las características a tratar en el trabajo y después de esto se delimita una muestra menor de la población, aproximadamente del 10% para que sea representativa, es decir, que contenga las características a analizar, y posteriormente hacer la transcripción, que toma por lo menos diez horas por cada hora de audio; esto podría parecer complicado, sin

embargo, el uso lingüístico suele ser más homogéneo que otras conductas humanas, como sostiene la autora.

El estudio del habla ha sido lento porque la estructura lingüística ha sido relacionada con la homogeneidad y porque el habla plantea varios problemas al estudiarla. Labov (1983) nos muestra algunos de ellos: presencia de agramaticalidad, variación en el habla y en la comunidad lingüística, la grabación y audio complicados y la rareza de las formas sintácticas. Este mismo autor sostiene que todos estos problemas no representan realmente un impedimento para no estudiar el habla, puesto que aproximadamente un 75% de los enunciados son frases formadas correctamente y si se hace a un lado titubeos y falsos comienzos, la cifra será mayor: "las variaciones en el lenguaje son el resultado de factores lingüísticos básicos". Por supuesto todo esto es más sencillo si se aprovechan las nuevas tecnologías.

¿Cómo deben ser considerados los análisis sociolingüísticos que se hacen en la actualidad? Moreno Fernández ha señalado que algunos autores dictan que como modelos idealizados de un fragmento de la lengua[5].

[5]"*Modelos*: se trata de artificios teóricos que pretenden esquematizar la organización interna de los hechos; *idealizados*: suponen un grado mayor o menor de abstracción, de despliegue de la realidad; *fragmento de una lengua*: sería muy difícil dar cuenta de todos los niveles lingüísticos y de los

La sociolingüística tiene un problema a la hora de utilizar grandes corpora como el CORDE (Corpus Diacrónico del Español) y el CREA (Corpus de la Referencia del Español Actual). La página oficial de este último corpus expone que éste tiene la intención de proporcionar "una muestra representativa y equilibrada del español estándar que se utiliza actualmente en el mundo". Para este corpus se tomó en cuenta el lapso de tiempo que comprende de 1975-1999, también se toman en cuenta textos de América[6] y españoles con un 50% cada uno y el 90% de esos textos proceden de textos escritos mientras que el 10% proviene de textos orales. El CREA y el CORDE resultan muy idealizados y son parciales, pues se estudian ciertos niveles lingüísticos y estilos sin ponerlos en relación con el resto de posibilidades en la lengua**.** En descargo, debe admitirse también que la idealización es inevitable porque no son los grupos sociales los que proporcionan los datos tangibles, sino

---

factores que en ellos se implican en un solo estudio y de forma exhaustiva." (Moreno Fernández, 1990:48, cursivas en el original).

[6] De este continente el CREA toma en cuenta seis zonas, la mexicana que comprende a México, el Sudeste de Estados Unidos, Guatemala, Honduras y El Salvador con el 40%, la zona andina que es parte de Venezuela, parte de Colombia, Ecuador, Perú y Bolivia con el 20%, la zona caribeña, Panamá, Rep. Dominicana, Costas de Venezuela y Colombia, Noroeste de Estados Unido, Cuba y Puerto Rico con un 17%, la zona rioplatense que se conforma por Argentina, Paraguay y Uruguay con el 14%, la zona chilena con el 6% y la zona central que corresponde a Nicaragua y Costa Rica con el 3%.

los individuos (paradoja saussureana)[7] y además, no se puede recoger datos de todos los hablantes a menos que el grupo de hablantes sea muy pequeño.

Kibrik señala que la investigación lingüística gira en torno a tres puntos: el investigador, la lengua (datos) y el modelo creado[8]. La relación entre estos puntos conduce a la idealización porque entre el investigador y los datos surge la paradoja del observador (mencionada arriba) y la *paradoja acumulativa* (cuanto más se sabe de una lengua más se puede saber de ella);entre los datos y el modelo, a través del investigador, se produce la paradoja saussureana; entre el modelo y el investigador nace un doble problema: mediante un método deductivo el modelo está idealizado; mediante un método inductivo lo está igualmente, puesto que no se parte de todos los datos.

[7]Citada en página 3.

[8]A. E. Kibrik, *The Methodology of Field Investigations in Linguistics,* citado en Moreno Fernández (1990: 49).

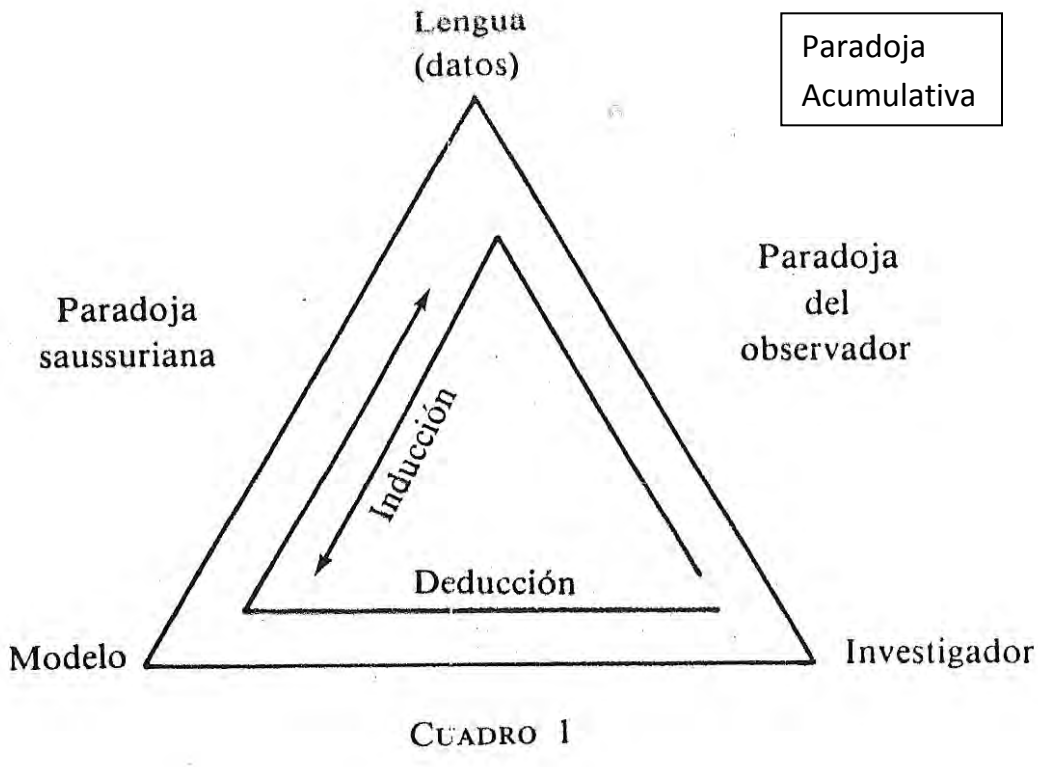



Fig. 3. *Esquema tomado de Moreno Fernández (1990)*[9].

1.5.1. *El corpus sociolingüístico y su representatividad*

Para formar un corpus que proporcione datos y resultados de un análisis lo más completo y realista posible no siempre es indispensable contar con todos los datos de un grupo, pero sí es necesario que contenga la mayor cantidad de cambios sociolingüísticos que se llevan a cabo en el grupo. Biber (1994) dice que la representatividad es "el grado en que una muestra incluye toda la gama de la

[9]*Nota: se agregó en el esquema la "paradoja acumulativa" porque Moreno Fernández habla de ella más no la registra en el esquema.*

variabilidad en la población", además de que la representatividad debe ser tomada desde el punto de vista de la situación y no sólo desde el criterio de la lingüística. Tognini-Bonelli (2001:55) refiriéndose a los corpora que se deben utilizar dice: "los textos son seleccionados de acuerdo con criterios explícitos, para captar las regularidades de una lengua, una variedad de la lengua o una lengua secundaria" (traducción mía).Además de esta definición existen otras menos completas, pero con las que también se puede estar de acuerdo[10].

Los corpora necesitan poseer naturalidad completa, el mejor corpus es el que se obtiene de lo vernáculo, de las espontaneidades, de los "errores" que se tienen al hablar y al escribir, se necesitan para un corpus de sociolingüista variacionista los cambios que se están llevando a cabo en el momento del habla cotidiana para poder captar las variaciones y reglas que posee la lengua: "todo el material incluido en un corpus, ya sea hablado, escrito o recogido a lo largo de cualquier dimensión intermedia (Biber 1998), se supone que es tomado de la

[10] P. ej.: "un corpus es una colección de textos que se supone representativa de una determinada lengua, dialecto, u otro subconjunto de un idioma que se utilizará para el análisis lingüístico" (N. Francis, 1992:17); o: "un corpus es una colección de texto de origen natural en el idioma, elegido para caracterizar a un estado o la variedad de un idioma" (Sinclair 1991:171).

comunicación genuina de la gente haciendo su trabajo normal" (Tognini-Bonelli 2001).

Como ya se explicó la representatividad en un corpus es primordial para su validación pues abarca el aspecto de las variaciones en la lengua, con un corpus representativo se puede trabajar, aunque existe también otro problema que es la posibilidad de que el lingüista no entre en contacto con los diversos contextos comunicativos, lo que empobrecería la calidad del corpus.

Moreno Fernández (1990) explica que existen dos tendencias para estudiar los datos, la primera es la que apoya Labov y es la que se centra en datos representativos con un análisis cuantitativo. La segunda tendencia es la que se centra en comunidades reducidas con una perspectiva cualitativa.

Debemos buscar que las conclusiones sociolingüísticas sean menos idealistas y trabajar en proporcionar datos reales para fraguar conclusiones *ad hoc* con la realidad. La sociolingüística se ha podido acercar a lo real por la naturaleza de los datos recogidos, pero aún existe el problema de que esos datos no sean representativos pues no se puede recoger la totalidad de los datos (por gasto de dinero y de tiempo).

## 1.6. RECAPITULACIÓN

Estas reflexiones obligan a pensar en la mejor forma de obtener el dato en sociolingüística, que es el objetivo principal de esa tesis. El mejor dato de la lengua es el dato vernáculo pues, como ya se comentó, es el dato que se da en el momento en el que se ejerce la lengua más cotidiana e íntima y el que proporcionará información certera y real. Esta será de gran apoyo para poder definir y delimitar lo que es la lengua que utilizamos.

En general, la orientación que se le ha dado al estudio de la lengua no ha contado con un enfoque en el que el contexto sea tomado en cuenta para dar una explicación más realista del fenómeno social del lenguaje. Sin embargo, desde hace más de medio siglo se ha intentado aclarar la importancia que tiene la estructura social sobre el cambio y la variación en la lengua. En esta tesis se expondrá una metodología y sus resultados como propuesta para un recogimiento de datos vernáculos que ayuden a la creación de corpora más realistas.

# CAPÍTULO 2

# MÉTODOS

La metodología propuesta busca obtener una muestra de lengua lo más cercana a la realidad que evite modificaciones forzadas por parte del hablante. Esta es la primera vez que se aplica en México una metodología de seguimiento de un informante a lo largo de un día normal, y como todo trabajo innovador,se fue enriqueciendo en la medida en que fue aplicada.

## 2.1. INFORMANTES

En un principio se decidió seleccionar a dos personas universitarias (una mujer y un hombre) que fueran cercanas a la investigadora para causar los menores problemas posibles y compartir el tiempo de todo un día con cada uno de ellos. Se planeó que el proyecto comenzara a las 8:00 am y concluyera a las 10:00 pm con el propósito de abarcar la mayor cantidad de horas en la lectura y habla de un hablante universitario común durante un día en la Ciudad de México. También se proyectó que se haría un seguimiento de cerca a los informantes a donde quiera que fueran y con una grabadora pequeña se grabaría discretamente no sólo lo que ellos dijeran, sino también lo que dijeran sus interlocutores, por lo cual habría sido necesario tener cuidado para que estos últimos no percibieran la

grabación[1]. Posteriormente los informantes indicarían a la investigadora la relación que tienen con cada interlocutor de ese día. Además del registro de habla, se tenía planeado tomar registro de la lectura y escritura de los informantes, para lo cual se tomarían fotos o fotocopias de todo lo que hubieran leído o escrito.

Se pensó en dos informantes para tener un registro de ambos sexos y de la edad promedio en la ciudad de México,que en el 2010 es de 31 años, según datos del INEGI, además de recopilar información de un sector particular de la población: el universitario promedio.

Esta propuesta se piloteó con un informante de manera parcial, perohubo complicaciones que impidieron su viabilidad:

- Encontrar a un informante. Aunque fueran amigos o parientes de la investigadora,los "candidatos" argumentaban que se iban a sentir incómodos con una persona grabando a su lado durante todo el día.
- Grabar durante todo el día al informante. Una vez que se obtuvo la aceptación y participación de un informante para una prueba, éste no logró mantenerse tranquilo en muchas situaciones como cuando deseaba

[1] Después de grabar se le informaría al participante grabado que se realizó una grabación secreta y se le pediría su autorización para usar su participación en la tesis.

hablar de temas personales o hablaba con otra persona, ya que le pidió a la investigadora en repetidas ocasiones que no grabara esos momentos: pedía privacidad.

- Mantener la grabadora encendida en todo momento. Al desconocer la investigadora el momento en el que el informante hablaría, debía mantener la grabadora encendida casi en todo momento lo que, por cuestiones técnicas, era difícil de lograr.
- Registrar la totalidad de lo que los informantes dijeron y/o les dijeron. Por cuestiones de privacidad,el informante de la prueba no permitió grabar algunos momentos.
- Mantener la distancia entre la investigadora y los informantes. Al tratarse de personas cercanas a la investigadora, con frecuencia se alteraba el contexto comunicativo por la cercanía entre ellos.
- Conservar la naturalidad. Al saberque estaba siendo grabado,elinformante perdía la naturalidad al hablar y además tenía que atender al registro de todas sus lecturas, pues en este método los informantes tienen que indicar qué es lo que leen y cuándo leen, para poder registrarlo; esto confundía al informante de la prueba y lo hacía sentirse incómodo, incluso argumentó que su lectura no era natural, es decir, no leía lo que en una circunstancia "normal" leería.

Para poder llevar a cabo el proyecto, la investigadora optó por elegir sólo un informante, entonces se propuso que fuera la misma investigadora y, aunque se regularon temas íntimos, en la auto-grabación se obtuvieron mejores resultados que con el informantede la prueba piloto. Las ventajas de que la investigadora sea la informante principal fueron las siguientes:

- La informante. Al tratarse de la investigadora, no hubo incomodidad por tener a una persona con una grabadora registrando en todo momento.
- Grabar durante todo el día al informante. Al tratarse de la investigadora, la grabación se hizo en el tiempo establecido.
- Mantener la grabadora encendida. Las grabaciones que se crearon fueron precisas, porque aunque la grabadora no pudiese mantenerse encendida todo el día, se podía saber en qué momento se hablaría o le hablarían ala informante.
- Registrar la totalidad de lo que los informantes dijeron y/o les dijeron. Al tratarse de una grabación secreta, los interlocutores de la informante expresaron absolutamente todo lo que quisieron con

naturalidad, es decir, no fingieron ni cuidaron lo que dijeron, aún cuando se tratara de asuntos privados.

- Mantener la distancia entre la investigadora y los informantes. No se dio este problema.
- Conservar la naturalidad.
    a) Los registros de lo que se leyó no fueron forzados y no hubo tantos problemas en cuanto a la naturalidad de lo que se dijo y lo que se leyó y registró.
    b) La lengua fue más detallada, pues al no sentirse observada por nadie, incluso la informante olvidaba que estaba siendo grabada.
    c) Hubo más naturalidad, y si bien la lengua no fue completamente vernácula, pues se tenía conciencia de la grabación, eso no impidió que se hicieran monólogos e incluso se cantara en la grabación, sin ningún obstáculo por vergüenza o por sentirse observado.

Este estudio no pretende expresar resultados completamente definitivos e indiscutibles, el estudio con un solo individuo no es necesariamente representativo, es sólo un intento y ensayo hacia un acercamiento a lo real, para

encontrar lo representativo sería necesario hacer varios estudios de informantes de diferente edad, nivel de educación y sexo para obtener una media con todos los datos. La informante no pertenece al promedio de las personas que viven en la Ciudad de México, pues como se mencionó, el promedio de edad es de31 años según cifras del INEGI y la investigadora contaba con 24 años el día de la grabación; el promedio de escolaridad en la Ciudad de México según también el INEGI es segundo grado de la educación media superior, mientras que la informante cuenta con un grado de licenciatura y forma parte de una segunda generación de profesionistas, además de que es egresada de la carrera de Letras y Literaturas Hispánicas, lo que indica un evidente gusto por la lectura y la escritura y contiene un vocabulario posiblemente más nutrido que el mexicano promedio. Lo anterior demuestra que la informante no forma parte del promedio de mexicanos aunque sí se encuentra dentro del 49.7% del total de la población que abarca de los 15-64 años (INEGI)y ciertamente, al tratarse de una institución pública, puede ser representativa hasta cierto punto de la población estudiantil femenina universitaria en la ciudad de México. En todo caso, un estudio de esta naturaleza, como se verá, permite una mayor profundidad analítica en términos cualitativos.

## 2.2. Proceso de grabación y registro de los datos

La grabación se hizo con una grabadora de voz de un reproductor de mp3 *Sony Walkman NWZ-B152F* de 2 Gb. Se grabó en formato *wav* a 16 bits. La nitidez de grabación para la investigación, aunque es muy buena para los sonidos que se captan en una distancia corta, tiene problemas en cuanto a las distancias largas; en cuanto al tiempo de grabación, puedemantenerse grabando hasta 60 minutos, transcurridos estos se debe programar la grabadora de nuevo para que comience otra grabación; por esta razón la informante debía estar al pendiente del tiempo para no perder detalle de lo que le pudiera oír y decir. El dispositivo de grabación por su tamaño y por su aspecto facilitó la grabación, no era evidente que grababa en secreto, por lo que se podía obtener mayor calidad en el sonido.

A continuación se presenta una bitácora que contiene los eventos comunicativos que se presentaron en un día común de la informante, ésta tiene como fin sólo indicar los eventos, no se trata de un estudio de etnografía del habla (Hymes 1974), es sólo un esquema de las interacciones comunicativas que se dieron durante el día.

Cuadro 2.1. *Bitácora de la grabación*

(Nota: los horarios anotados son aproximados)

| **Hora** | **Evento comunicativo y lugar** | **Duración aproximada** |
|---|---|---|

| | | |
|---|---|---|
| 8:00 | INTERACCIÓN ORAL. SALUDOS<br>Saludo a mi abuela S y a la Sra. C. Interacción muy breve. | 30 seg. |
| 8:30 | CIBER-INTERACCIÓN.<br>Diálogos breves con mi hermana P y con mi amigo U. | 20 min. |
| 10:00 | Lectura de subtítulos por TV de una serie en CD. | 30 min. |
| 11:00 | Interacción por celular con mi amiga A. | 5 min. |
| 11:15 | Búsqueda de información por Internet y lectura de artículos e información. | 40 min |
| 12:00 | INTERACCIÓN ORAL. DESPEDIDA Y SALUDO<br>Despedida a mi abuela S. y a la Sra. C. Saludo al Sr. Don M. Interacción muy breve. | 1 min. |
| 12:15 | INTERACCIÓN ORAL<br>Llegada al gimnasio.<br>Saludo con H y saludo y plática con M, trabajadora del gimnasio. | 3 min. |
| 12:30 | LECTURA DEL CELULAR | 30 seg. |
| 2:30 | INTERACCIÓN ORAL<br>Plática con H, trabajador del gimnasio<br>Salida del gimnasio. | 2:30 min. |
| 3:00 | INTERACCIÓN ORAL<br>Llegada a la casa.<br>Plática breve con C y plática con K. | 15 min. |
| 3:45 | INTERACCIÓN ORAL<br>Plática breve con S y C. | 1 min. |

| | | |
|---|---|---|
| 4:00 | INTERACCIÓN ORAL<br>Plática con K. | 2min. |
| 4:45 | INTERACCIÓN ORAL<br>Plática con K. | 7 min. |
| 5:30 | Llamado de E para K e interacción con E. | 1 min. |
| 5:31 | RADIO<br>Comenzó a escucharse la canción de *The Cure, "*Love song*"*. | 1 seg. |
| 5:32 | INTERACCIÓN ORAL<br>Plática con K. | 4 min. |
| 5:35 | LECTURA<br>Indicaciones de una cámara fotográfica. | 2:20 min. |
| 5:42 | INTERACCIÓN ORAL<br>Plática con E. | 1 min. |
| 5:46 | MONÓLOGO<br>Mientras lee indicaciones de la cámara. | 15 min |
| 6:30 | LECTURA<br>Lectura en internet. | 5 min. |
| 6:31 | LECTURA EN VOZ ALTA | 5 min. |
| 6:45 | LECTURA y ESCRITURA<br>Lectura en Internet.<br>Termina de escucharse el radio. | 5 min. |
| 6:50 | INTERACCIÓN ORAL<br>Plática con K. | 10min. |

| | | |
|---|---|---|
| 7:00 | LECTURA Y LECTURA EN VOZ ALTA<br>En internet. | 5min. |
| 7:30 | INTERACCIÓN ORAL<br>Plática con K. | 5min. |
| 7:50 | LECTURA<br>En Internet. | 5min. |
| 8:00 | INTERACCIÓN ORAL<br>Plática con K. | 10 min. |
| 8:30 | LECTURA<br>En Internet. | 2min. |
| 9:00 | LECTURA<br>Capítulo 8 del texto de W. Labov, *Modelos Sociolingüísticos.* | 7 min. |
| 9:30 | INTERACCIÓN ORAL<br>Plática con K. | 10 min. |
| 10:00 | LECTURA<br>Lectura de libros, discos y canales de TV. | 5 min. |

Para dar mayor detalle de las interacciones, se presentaa continuación un breve relato de los primeros eventos del día, donde se describe a detalle lo que se hizo durante ese lapso de tiempo.

El proyecto comenzó a las 8:00, la hora en la que la informante se levantó, un día antes se preparó la grabadora y una libreta para registrar rápidamente, para que fuera más sencillo y cuidado el seguimiento y no se perdiera ningún detalle.

Una vez despierta supo que no hablaría inmediatamente, por lo que no comenzó la grabación hasta que se encontró con otra persona con la que sabía que iba a interactuar verbalmente y se daría el buenos días, posteriormente, aunque sabía que muy probablemente no hablaría con nadie, sabía que podía hablar consigo misma. Leyó varias etiquetas de ropa y marcas de aparatos electrónicos, después llegó otro posible participante en un diálogo o simplemente otro participante con grandes posibilidades de hablar y ser escuchado por el informante por lo que se mantuvo prendida la grabadora durante diez minutos.

Posteriormente, se encendió la computadora y se leyó casi sin atención los anuncios de la preparación del equipo de computación y lo que ofrece el *messenger* –como opciones de aplicaciones y mensajes ya sean personales, de algún contacto o mensajes de actualizaciones de algún contacto. Cuando se apartó la vista del monitor, se vio la portada de un libro, posteriormente se siguió leyendo en la computadora, pues se accedió a una cuenta de correo electrónico para lo que se necesitó escribir la cuenta personal, posteriormente se leyeron las opciones que tiene este servidor, nombres de contactos y nombres de *mails* además de *Messenger* donde, si se usa, se leerá repetidamente el nombre con el que se está compartiendo un mensaje y el nombre propio, aunque no se leyó con atención sí se registró cada vez que se leyó un mensaje.

El primer contacto con la escritura es por medio de Internet y es para entablar un diálogo cibernético en escritura, posteriormente se entabló otra conversación por medio del *messenger*.

Hasta aquí los primeros minutos de la grabación. Este breve recuento permite al lector imaginar los distintos tipos de interacción verbal que se producen al inicio de un día común y de aquí la diversidad de tipos textuales que se detallan en el siguiente apartado.

## 2.3. CRITERIOS PARA LA TRASCRIPCIÓN DE LA LENGUA HABLADA EN EL CORPUS

La transcripción del corpus en audio, en la computadora y en una libreta se hizo con un método inspirado en el *Corpus Sociolinguístico de la Ciudad de México* (Martín Butragueño y Lastra 2011). Los detalles del sistema se presentan en el Apéndice I.

El registro del habla en transcripción y digitalización es de suma importancia para obtener a partir de éstos un análisis lingüístico y realista. Existen distintos programas que pueden utilizarse para facilitar la transcripción y digitalización de un corpus como *Toolbox* o *Simple Concordance Program* así como también existen modelos de transcripción con adaptación como TEI *(Text*

*Encoding Iniciative)* para proyectos muy grandes como CHILDES y PRESEEA (Proyecto para el Estudio Sociolingüístico del Español de España y América)[2].

## 2.4. PROCESAMIENTO DE DATOS

Para el análisis del léxico obtenido en las grabaciones y los textos escritos recogidos, se utilizó el programa *Ant Conc 3.2.2.1w* (Laurence 2010)[3]. El texto se introdujo en el programa para encontrar la cantidad de palabras gráficas (a las que llama *tokens*) y la cantidad de tipos de palabras gráficas llamadas *types*, estos conceptos no deben confundirse con lexemas, a continuación una tabla aclaratoria.

*Type* →

| *Palabra* | *Frecuencia* |
|---|---|
| *a* | 894 |
| *de* | 567 |
| *y* | 387 |

← Número de *tokens* del *type* "de".

[2] Información tomada de Serrano (2009).

[3] Programa computacional que puede ser descargado desde la siguiente página: http://www.antlab.sci.waseda.ac.jp/software.html y del que se incluye un ejemplo para mayor información de su funcionamiento en el Apéndice II.

Para cada división en la tabla de resultados se introdujo el texto sin formato para su fácil lectura y se indicó: a) que se tomaran en cuenta de igual manera las mayúsculas y las minúsculas; b) para el caso de las palabras en inglés, que se tomaran como palabras independientes aquellas que se escriben con abreviaturas como: *'re* (*are*); *'t* (*not*); c) eliminar las abreviaturas de los nombres de los participantes y d) no contar todo texto entre paréntesis. El programa *Ant Conc* sólo reconoce y cuenta las palabras como una secuencia de letras con espacios en blanco a la derecha y a la izquierda; para este proyecto no se contaron los signos de puntuación ni los númerossi estos fueron registrados en cifra. Para poder transcribir se utilizó el programa *Wave Pad Sound Editor* v 4.52[4]donde se borraron los silencios muy largos que se dieron en su mayoría entre los distintos eventos comunicativos y que no eran necesarios mantener en la grabación[5], se repitió el texto cuantas veces se quiso, también ayudó a seleccionar y modular el texto de las grabaciones para que fuera más fácil la transcripción. Aunque estos programas ayudaron en la recopilación de datos, evidentemente muchísimo trabajo para identificar los vocablos y lexemas correspondieron a la investigadora.

[4] Este programa puede ser descargado desde la página: http://download.cnet.com/WavePad-Sound-Editor/3000-2170_4-10276212.html. Su función es editar grabaciones, recortar, pegar, borrar, etc.

[5] Estos silencios no se pierden pues están registrados en el inter de los eventos comunicativos.

Dentro del corpus los textos leídos se escanearon, se copiaron de Internet o se copiaron de la libreta utilizada para el registro leído, los diálogos de la serie televisiva en disco se copiaron de Internet, lo escrito se copió de la computadora y celular; estos textos no se transcribieron con los criterios para los diálogos, pues no es posible integrarlos.

En la transcripción de esta investigación fue necesario utilizar un código de colores: azul para el texto leído en silencio, rosa para el texto escrito y verde para el texto leído en voz alta y para los textos de diálogos, monólogosy texto sólo escuchado no se utilizó ningún color.Este código se implementó para seguir fielmente el orden de lo que se hizo en un día.

Para observar la forma en que se organizó el corpus se incluyeron ejemplos de cada división donde se puede apreciar cómo se incluyeron los marcadores de colores, criterios de transcripción y *marginalia*, que se incluyeron en el conteo de frecuencia en *tokens* y *types*.

2.4.1. *Ejemplos de tipos textuales*

2.4.1.1. Diálogos

Y: tengo una pregunta/ ¿y el entrenador?

M: ¿Hugo?/¿Gus?

Y: eh:

M: Gustavo

Y: sí bueno/ el que [estaba <…>]

M: [ya no va a estar con nosotros]

Y: ¿ya no?//¡ ú:fale!<~úfale>

M: sí/ cerramos sí ¿sí has leído verdad?// que cerramos el día dos/ definitivamente

Y: ¿en serio?

M: sí

Y: ¿por qué?

M: por//hora sí que por exceso de// de clientela

Y: no:<~no>// ¿el dos de diciembre?

M: sí las personas que: este:// que hayan /cómo se llama/ pagado su osea sólo este mes está bien pero ya en enero no se recibirán más pagos <…>

2.4.1.2. Monólogos. El texto de los monólogos se constituye comúnmente de oraciones que produce la informante para la misma informante y explicaciones que comúnmente se daríansólo en el pensamiento. En este ejemplo de monólogo hay intercalaciones de texto leído en voz alta, que son las palabras sombreadas.

Y: (para sí misma)esta cosa…(leyendo para sí misma) ajustes uno // ¡muy bien!// ¡bien!/ aho:ra:<~ahora>// no a ver/ pero por qué está en *off*// modo video:/ ¡vídeo! ///¿por qué no? (silbido bostezo ) m mm m:… optimiza automáticamente los ajustes de la cámara atendiendo al tipo de disparo/ para fotos de documentos etcétera// a ver vamos a tomar una foto tipo retrato ah/¡esa cosa de allá!//¡no estuvo tan horrible!/ ¡ahora a esto! /ah no pero hay un especia:l<~especial>/ para:…cosas de la *web*// m: no/ quién sabe/ me perdí/// ash/ /bueno ya (silencio) (silbido) (bostezo) ay: <~ay>¡qué horror!

2.4.1.3. Texto escrito. En este ejemplo de texto escrito se observa un diálogo vía escritura, por esa razón es que la parte de escritura está sombreada en rosa y la de lectura en silencio en azul.

Penélope: bien!!!!!!!!!!!!!!!

perfecto que estesaqui!

yo:pepe

Penélope: no recuerdo el telefono de charlie!

yo:estoy viendo que me mandaste unas fotos

te lo paso

Penélope: gracias!

yo:yy no me andste nunca tus otras fotos

Penélope: cuales?

yo: pues este mail que mandaste ayer dice

que estás son las últimas de Rio

2.4.1.4. Texto leído en silencio. En este ejemplo se encuentra un texto como tal y palabras que fueron leídas por separado, en su mayoría palabras en inglés que corresponden a marcas comerciales.

Reebook

Adidas

Reebook

Women secret

East pak

Kyoto

Sony

**El filósofo que vino del frío**

SlavojZizek nació en 1949 en Eslovenia, en la antigua Yugoslavia. Quería ser director de cine, pero terminó siendo filósofo. En los últimos años del comunismo participó en los movimientos

democráticos de oposición al régimen y en 1990 con las primeras elecciones libres fue candidato a la presidencia de la República Eslovena por una amplia coalición de izquierdas.

2.4.1.5. Texto leído en voz alta. Este texto fue leído de Internet en *Wikipedia* y es un texto que contiene información científica.

Un electrolito// es cualquier sustancia que contiene iones libres/ los que se comportan como un medio de/ como medio conductor eléctrico/ m: debido a que generalmente consisten de iones en solución/ los electrólitos/ también son conocidos como/soluciones iónicas// pero también son posibles electrolitos fundidos y electrolitos sólidos//

2.4.1.6. Texto híbrido. Este texto fue recibido tanto de forma escrita como oral a través de un programa de TV que contenía audio en inglés y subtítulos en español.

Por otra parte,
algunos físicos están preocupados
de que si este supercolisionador
en efecto funciona,

creará un hoyo negro

que absorberá la Tierra,

2.4.1.7. Música. En su mayor parte las canciones escuchadas se encuentran en inglés, esto se debe a la influencia de este idioma en nuestro país, además de que la informante tiene una inclinación por música en inglés.

*It´s too late to change events*

*It´s time to face the consequence*

*For delivering the proof*

*In the policy of truth*

*solo ve como me quedo aquí esperando a que no estés*

*en espera de que vuelvas, y tal vez vuelvas por mí*

2.4.1.8. Ruido de fondo. En su mayor parte es texto de anuncios de radio.

(Anuncio radiofónico) A dónde va con <…> vecina voy a venderla para comprar medicinas <…> métela al seguro popular <…> ánimo te acompaño/ durante el gobierno del presidente de la república más de cuarenta y un millones de mexicanos ya cuentan con seguro popular…

## 2.5. CONCLUSIONES

Con la metodología propuesta se obtuvieron datos de habla espontánea que son en los que la sociolingüística variacionista se interesa y por ende los que estábamos buscando para la creación de un corpus socialmente realista. La forma de llevar a cabo el proyecto fue la menos problemática ya que la investigadora se comportó lo más realista posible ante la grabación y las personas que interactuaron con ella fueron completamente naturales pues no supieron de esta recolección de información realizada en secreto. La información, junto con los nombres de los participantes, se mantuvieron en secreto para no tener problemas éticos. La inclusión a detalle de una parte de la bitácora de eventos comunicativos y los fragmentos transcritos de cada tipo textual sin modificar su entorno dentro del orden del corpus, proporcionan un acercamiento minucioso al ensayo del método que podría servir para recolectar información y hacer corpora socialmente realista y análisis de la misma naturaleza.

# CAPÍTULO 3

# UN ANÁLISIS DEL LÉXICO EN DIVERSOS TIPOS TEXTUALES

## 3.1. INTRODUCCIÓN

Esta parte de análisis corresponde sólo al léxico, no se abordan otros niveles de la lengua como el morfosintáctico o fonético. El tema central por lo tanto es analizar básicamente el volumen y características del léxico en los diferentes tipos textuales y no las formas particulares de comunicación entre los informantes involucrados.

Los resultados del registro de la recopilación de vocablos indica un total de 19,989 *tokens* (palabras) y 4,178 *types* (tipos de palabras o vocablos); para obtener estos resultados se tomaron en cuenta todos los medios por lo que se considera que la lengua se expresa, es decir, oralidad, escritura y lectura. Dentro de estos medios encontramos todo lo que expresamos con la lengua (lo que hablamos y escribimos o *output*) y todo lo que percibimos con ella (lo que oímos y leemos o *input*). Se decidió hacer esta división pues consideramos que de esta forma habrá mayor claridad en los resultados y se podrá llegar a una conclusión con mayor detalle. La separación en tipos textuales se dio en: Oralidad –aquello que se expresó y recibió mediante el sonido-, Escritura-todo lo que se escribió-, Lectura en silencio, Lectura en voz alta, Híbridos -aquello que mezcló dos tipos textuales, en

este caso fue una mezcla entre Oralidad y Lectura-; se decidió incluir un apartado para Música -todo aquello que se escuchó y cantó con una letra y tonada determinadas; y finalmente un tipo que se llamará aquí Ruido de fondo, esto es, todas las expresiones lingüísticas que aparecen en las grabaciones pero a las que no se presta atención (mensajes comerciales, pláticas ajenas y anuncios en video).

### 3.2. ORALIDAD

Dentro del tipo textual de oralidad se encuentra el **Diálogo**[1] que suma un 28.3% del total del corpus, esto indica que poco menos de una tercera parte del total de **Input** y **Output** en un día es interacción oral. Tomando en cuenta el texto de diálogo como el 100%, éste se divide en 49.4% de input 50.5% de output, lo que indica una relación bastante equilibrada entre lo que se dice y lo que se escucha.

[1] El *Diccionario básico de lingüística* (Luna Traill, Vigueras & Baez 2007) define el diálogo como: "Proceso interactivo de la comunicación formado por una cadena de intervenciones lingüísticas y que se desarrolla con los interlocutores cara a cara" (p. 446).

Tabla 3.1. *Diálogo en el corpus*

| Participante | Tokens | Types | % Diálogo | % en Corpus total |
|---|---|---|---|---|
| *Yael* | 2864 | 701 | 50.5% | 14.3% |
| *Otros* | 2803 | 737 | 49.4% | 14.0% |
| *Total* | 5667 | 1099 | 100% | 28.3% |


Figura 3.1. *Participantes en el diálogo*

La interacción persona a persona es equitativa, está equilibrada e incluso dentro del corpus podemos encontrar que se respeta el turno de cada participante, hay muy pocos traslapes pues se intenta mantener un espacio de tiempo para cada hablante, quienes tienden a respetar el turno.

El **Monólogo** ocupa sólo un 2% del corpus total, este porcentaje representa las palabras que, a diferencia del diálogo, no fueron planificadas por la informante,

fueron espontáneas; en esta parte, aunque la informante *no comunica* estrictamente información a otra persona, se está reafirmando algún pensamiento que para ella no contiene una importancia relevante como para expresarlo a otro individuo y crear un diálogo, pero sí la suficiente importancia como para que el pensamiento sea expuesto para el mismo informante y sea una reafirmación del pensamiento en sí (Silva- Corvalán: 2001, 24).

En resumen, la **Oralidad** posee el porcentaje más alto dentro de la tabla con un 30.4%, dejando así un 69.6% para el resto del corpus que contiene Escritura, Lectura en silencio, Lectura en voz alta, Híbridos, Música y Ruido de fondo. Cada sección contiene una densidad léxica, esto es, lo que se obtiene de la división de total de *tokens* (5667) entre el total de *types* (1099), en esta sección la densidad léxica es de .19, esto significa que por cada 100 palabras hay 19 vocablos distintos.

Tabla 3.2. *Peso de la oralidad frente a los demás tipos textuales*

| **Tipo textual** | **Total Tokens** | **Peso % en corpus (en cuanto a tokens)** |
|---|---|---|
| *Oralidad* | 6086[2] | **30.4%** |
| *Resto del corpus* | 13903 | **69.6%** |
| *Total* | 19989 | **100%** |

[2] Se incluye el 2% de Monólogo

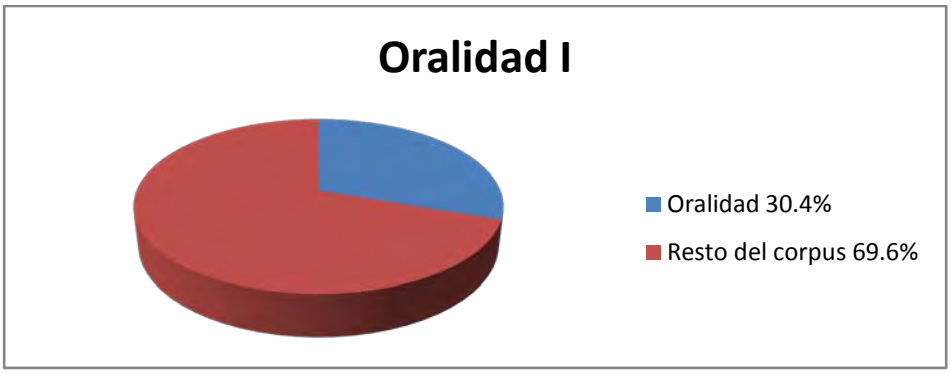



Figura 3.2. *Oralidad* vs. *resto de tipos textuales*

A continuación se presenta una tabla donde se muestran los *types* con mayor frecuencia en la sección de Oralidad, estas palabras son gramaticales, sin contenido léxico, y funcionan para dar coherencia a la lengua, palabras que en su mayoría son conectores, conjunciones, adverbios, un pronombre y un verbo copulativo.

Tabla 3.3.Types *con mayor frecuencia en Oralidad*

| Oralidad | | |
|---|---|---|
| Lugar | Frecuencia de aparición | Palabra |
| 1° | 291 | no |
| 2° | 192 | que |
| 3° | 182 | y |
| 4° | 173 | a |
| 5° | 149 | es |
| 6° | 148 | de |
| 7° | 132 | qué |
| 8° | 115 | la |
| 9° | 101 | sí |
| 10° | 98 | el |

Existe una división en semántica entre palabras con contenido léxico y palabras sin contenido léxico. Henry Sweet en su texto *New Inglish Grammar*(1898) y Stephen Ullmann (1972)en su texto *Semántica* entre otros hablan acerca de esta división. Las palabras con contenido léxico tienen sentido en sí mismas y casi siempre tienen un referente físico o imaginario y las palabras sin contenido léxico funcionan para dar coherencia a un texto y tienen sentido únicamente en contexto como los pronombres personales; Sweet y Ullmann las llaman *palabras plenas* vs. *Palabras formas*: las primeras, dice Ullmann (1972), son palabras como **árbol, cantar,**

**azul** y **suavemente**, éstas, aunque aparezcan aisladas, conservan algún significado,[3]mientras que las palabras plenas como: **el, la, lo, los, y, ello** y **de** no lo conservan[4], éstas mantienen una colaboración con otras palabras para mantener la coherencia, por sí mismas no mantienen ningún significado.

Tomando en cuenta lo anterior, además presentar tablas donde se encuentren los *types* más frecuentes, que comúnmente contienen palabras plenas, también se presentarán tablas con las palabras-forma más frecuentes.

La primera palabra con contenido léxico o palabra-forma, dentro de la división Oralidad, se situó en el puesto treinta y siete, esto es, los primeros puestos desde el uno hasta el treinta y seis fueron ocupados por palabras sin contenido léxico. Las tablas no contienen los verbos copulativos ni auxiliares, tampoco palabras que sean frases hechas como "bueno" refiriéndose a "está bien" y no al adjetivo ni tampoco nombres propios de personas. Esta tabla contiene dos nombres propios, cuatro verbos, tres sustantivos y un adjetivo calificativo.

Las primeras palabras con contenido léxico dentro de la sección Oralidad son las siguientes:

[3] Éstas suelen ser sustantivos, adjetivos y verbos.

[4] Suelen ser adverbios, pronombres y conjunciones.

Tabla 3.5. *Palabras con contenido léxico Oralidad*

| Lugar | Lugar en Tabla | Frecuencia de aparición | Palabra con significado léxico |
|---|---|---|---|
| 1° | 37° | 32 | decir |
| 2° | 50° | 22 | saber |
| 3° | 62° | 16 | cosas |
| 4° | 72° | 14 | *Slytherin* |
| 5° | 74° | 13 | gimnasio |
| 6° | 77° | 12 | creer |
| 7° | 81° | 12 | *Gryffindor* |
| 8° | 82° | 12 | gusta |
| 9° | 84° | 12 | horrible |
| 10° | 89° | 10 | amiga |

La sección **Oralidad**, como ya se comentó, puede ser dividida en dos bloques: **Diálogo** y **Monólogo**[5], el primero consta de un 93.1% y el monólogo con un 6.8%, este porcentaje, aunque no muy alto, no existiría o sería casi nulo en muchos otros casos.

[5] Posteriormente se mencionará un tipo textual que puede caber dentro de la Oralidad.

Tabla 3.5. *Comparación entre los datos de Diálogo* vs. *Monólogo*

| Oralidad | Tokens | Types | % Corpus | % Oralidad |
|---|---|---|---|---|
| *Diálogo* | 5667 | 1099 | 28.3% | 93.1% |
| *Monólogo* | 419 | 178 | 2% | 6.8% |


Figura 3.3. *Diálogo* vs. *Monólogo*

En estos resultados también podemos cuantificar el total de tipos de palabras o *types* que es 1099, dentro de estos encontramos que la informante-investigadora mencionó 704 tipos de palabras y sus interlocutores 738, casi la misma cantidad; además de este parecido encontramos que los tipos de palabras que más se encuentran son casi los mismos, lo que demuestra que el léxico es muy parecido en cuanto a cualidad y cantidad. A continuación unas listas de los *types* que tuvieron mayor frecuencia de aparición:

Tabla 3.6.Types *con mayor frecuencia en Diálogo Yael*

| Yael | | |
|---|---|---|
| Lugar | Frecuencia de aparición | Palabra |
| 1° | 138 | no |
| 2° | 95 | que |
| 3° | 90 | y |
| 4° | 78 | qué |
| 5° | 76 | es |
| 6° | 70 | a |
| 7° | 64 | de |
| 8° | 58 | el |
| 9° | 46 | la |
| 10° | 46 | pero |

Tabla 3.7.Types *con mayor frecuencia en Diálogo Otros*

| Otros | | |
|---|---|---|
| **Lugar** | **Frecuencia de aparición** | **Palabra** |
| 1° | 129 | no[6] |
| 2° | 96 | que |
| 3° | 92 | a |
| 4° | 90 | y |
| 5° | 77 | de |
| 6° | 68 | la |
| 7° | 65 | es |
| 8° | 55 | sí |
| 9° | 42 | pero |
| 10° | 38 | el |

Las palabras que se repiten más tanto en el diálogo Yael como en el **Diálogo Otros** es "no" seguido de "que", las siguientes palabras, aunque en distinto orden, son las mismas con excepción de "qué" en Yael y "sí" en Otros.

A continuación las tablas de contenido léxico:

[6] Muchos de estos "no" son en realidad marcadores discursivos como por ejemplo: "se ve bien ¿no?".

Tabla 3.8. *Palabras con contenido léxico Diálogo Yael*

| Lugar | Lugar en Tabla | Frecuencia de aparición | Palabra con contenido léxico |
|---|---|---|---|
| 1° | 14° | 33 | decir |
| 2° | 31° | 20 | saber |
| 3° | 49° | 13 | cosa(s) |
| 4° | 50° | 10 | creer |
| 5° | 69° | 8 | gimnasio |
| 6° | 79° | 7 | taxi |
| 7° | 81° | 6 | amiga |
| 8° | 82° | 6 | Dios |
| 9° | 84° | 6 | horrible |
| 10° | 90° | 5 | buena |

Tabla 3.9. *Palabras con contenido léxico Diálogo Otros*

| Lugar | Lugar en Tabla | Frecuencia de aparición | Palabra con contenido léxico |
|---|---|---|---|
| 1° | 11° | 36 | decir |
| 2° | 40° | 14 | saber |
| 3° | 43 | 12 | gustar |
| 4° | 54° | 9 | *Shlyterin* |
| 5° | 59° | 8 | *Gryffindor* |
| 6° | 66° | 7 | niña |
| 7° | 78° | 6 | examen |
| 8° | 82° | 6 | maestra |
| 9° | 91° | 5 | gimnasio |
| 10° | 99° | 4 | amiga |

Las tablas de diálogo muestran que la primer palabra con contenido léxico fue "dijo" en ambos casos en los puestos 34 y 35 del total de *types* con 16 apariciones en cada sección, otras palabras en común entre las tablas son "gimnasio", "amiga" "sé" lo que muestra que se habló de un mismo tema durante un lapso de tiempo considerable. La densidad léxica de esta sección es de .19 al igual que en Oralidad.

## 3.3. ESCRITURA

Con respecto a la **Escritura,** encontramos que en el total del corpus consta de un 1.6%, que corresponde a 322 palabras, ocho de las cuales estaban en inglés. La escritura se realizó en la computadora y en el celular y mucho de lo escrito fueron mensajes para otra persona en línea, fueron partes de conversaciones escritas que se dieron como se dan las conversaciones cara a cara, pues las respuestas se daban y recibían inmediatamente, también se escribieron mensajes que no se recibían inmediatamente, la otra parte del texto escrito fueron nombres para buscar en la red, muchos de estos fueron en inglés. La encuesta que elaboró CONACULTA en el 2010 proporcionó una cifra reveladora en cuanto a la escritura, pues expuso que el 8.6% de los encuestados escriben literatura en su tiempo libre, esto puede indicar que el porcentaje de individuos que escriben por trabajo o para comunicarse es mucho mayor.

En un día común la informante escribió apenas 322 palabras y196 *types.* No escribió *mails* ni cartas ni trabajos escolares ni por recreación, aunque sí lo suele hacer, es decir, fue un día de casi nula actividad en la escritura.

Tabla 3.10 *Datos de Escritura*

| TIPO TEXTUAL | Total Tokens | Total Types | Peso % en corpus EN CUANTO A TOKENS |
|---|---|---|---|
| *ESCRITURA* | 322 | 196 | **1.6%** |


Figura 3.4. *Escritura* vs. *Resto del corpus*

A continuación una tabla con los primeros puestos de palabras en frecuencia dentro de **Escritura.**

Tabla 3.11. Types *con mayor frecuencia en Escritura*

| Lugar | Frecuencia de aparición | Palabra |
|---|---|---|
| 1° | 16 | que |
| 2° | 12 | no |
| 3° | 10 | me |
| 4° | 9 | a |
| 5° | 9 | y |
| 6° | 8 | la |
| 7° | 7 | el |
| 8° | 7 | pues |
| 9° | 6 | de |
| 10° | 6 | qué |

La lista de palabras más frecuentes en **Escritura** es muy parecida a la que se encuentra en la tabla de **Diálogo**, esto demuestra que tanto en lengua escrita como en la lengua hablada la informante utilizó con mayor frecuencia palabras sin contenido léxico.

Ahora se muestra la tabla con las palabras con significado léxico o palabras llenas con más frecuencia en la Escritura.

Tabla 3.12. *Palabras con contenido léxico con mayor frecuencia*

| Lugar | Lugar en Tabla | Frecuencia de aparición | Palabra con significado léxico |
|---|---|---|---|
| 1° | 12° | 5 | dice |
| 2° | 24° | 2 | fotos |
| 3° | 29° | 2 | mandaste |
| 4° | 32° | 2 | ocupado |
| 5° | 34° | 2 | pensé |
| 6° | 37° | 2 | puntada |
| 7° | 45° | 1 | amibas |
| 8° | 48° | 1 | atrofiado |
| 9° | 56° | 1 | cabeza |
| 10° | 57° | 1 | calor |

La primera palabra, con contenido léxico, al igual que en Oralidad es "dice", y se encuentra en el número veintidós del total de *types*

La densidad léxica de Escritura es de.60, ésta es mucho más alta que la densidad de Oralidad.

## 3.4. HÍBRIDOS

Se creó un apartado de Tipo Textual llamado "**Híbridos**", esto es actividades en las que se combinan **Lectura** y **Oralidad**, específicamente en las películas y series de

televisión con subtitulaje. Un sector muy alto de la población --más del 40% de los hogares en México-- cuenta con televisión de paga, según LAMAC[7]; en la programación usual de TV de paga encontramos programas de habla inglesa con subtítulos en español, además de esta circunstancia en la que se presenta este híbrido de lectura en español y audio en inglés, encontramos películas o incluso las mismas series de TV en DVD. Este último caso fue el que se registró en el corpus con la serie *The Big Bang Theory* que la informante vio. Este apartado obtuvo un 11.8% del total del corpus; dicho porcentaje demuestra que este tipo de situaciones abarca un peso considerable en cuanto a la recepción de tipos textuales complejos que no están bien definidos y que por ello aquí llamamos **Híbridos**.

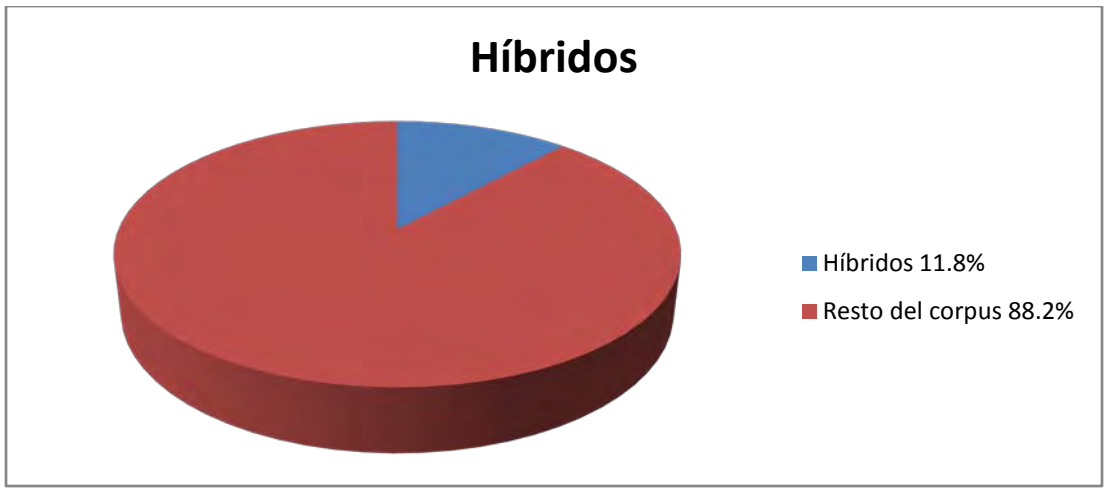



Figura 3.5. *Híbridos* vs. *Resto del corpus*

[7] http://www.lamac.org/mexico/metricas/total-por-tv-paga consultada el 23 de agosto 2011.

La sección de **Híbridos** constó de 2377 palabras o *tokens* y 927 tipos de palabras o *types,* de las que se registró una tabla de *types* con mayor frecuencia que a continuación se muestra:

Tabla 3.13. Types *con mayor frecuencia de aparición en Híbridos*

| **Lugar** | **Frecuencia de aparición** | **Palabra** |
|---|---|---|
| 1° | 95 | de |
| 2° | 77 | que |
| 3° | 53 | a |
| 4° | 51 | la |
| 5° | 51 | no |
| 6° | 41 | el |
| 7° | 37 | en |
| 8° | 35 | es |
| 9° | 33 | por |
| 10° | 31 | un |

Al igual que la **Oralidad** y la **Escritura**, la tabla de *types* de **Híbridos** contiene pocas palabras con significado léxico, el resto son palabras que mantienen la coherencia en un texto. Esto demuestra que el programa que se vio y leyó contiene también en su mayor parte palabras que conectan ideas y mantienen la coherencia. La lista de palabras con significado léxico en esta sección está encabezada por "hermana" ocupando el lugar veintidós dentro del total de *types*

con una frecuencia de quince veces registrada. La densidad léxica de Híbridos es de .38.

Tabla 3.14. *Palabras con significado léxico en textos Híbridos*

| Lugar | Lugar en Tabla | Frecuencia de aparición | Palabra con significado léxico |
|---|---|---|---|
| 1° | 17° | 20 | hermana (o) |
| 2° | 25 | 13 | querer |
| 3° | 38° | 9 | creer |
| 4° | 48° | 7 | amigo |
| 5° | 50° | 7 | decir |
| 6° | 65° | 7 | hablar |
| 7° | 66° | 6 | hombre |
| 8° | 87° | 4 | dios |
| 9° | 92° | 4 | espacial |
| 10° | 95° | 4 | habitación |

## 3.5. LECTURA

A diferencia de la **Escritura**, la **Lectura** obtuvo un porcentaje muy alto, con un 30.4% si juntamos **Lectura en voz alta**2.3% y **Lectura en silencio** 28.1%, y aún más si juntamos este porcentaje con lo que se encuentra en el tipo textual

**Híbridos**11.8%, en ese caso tendríamos un 42.2%, poco menos de la mitad del total del corpus, fue el porcentaje más alto.

La **Lectura en silencio** obtuvo un 28.1%, casi una tercera parte del total del corpus, en este apartado es en donde vemos un gran número de *tokens* con 5619 y el mayor número de *types,* 2043, esto se debe a que además de la lectura de las respuestas en las conversaciones escritas, mensajes, marcas y anuncios encontramos artículos y texto de un libro.

La densidad léxica de esta sección es de .36, de cada 100 palabras hay 36 vocablos distintos. En seguida se muestra la tabla de los *types* con mayor frecuencia en la Lectura en silencio.

Tabla 3.15. Types *con mayor frecuencia de aparición en Lectura en silencio*

| Lugar | Frecuencia de aparición | Palabra |
|---|---|---|
| 1° | 95 | de |
| 2° | 77 | que |
| 3° | 53 | a |
| 4° | 51 | la |
| 5° | 51 | no |
| 6° | 41 | el |
| 7° | 37 | en |
| 8° | 35 | es |
| 9° | 33 | por |
| 10° | 31 | un |

Tabla 3.16.*Lectura en silencio y en voz alta*

| | | | |
|---|---|---|---|
| **LECTURA SILENCIO** | 5619 | 2043 | **28.1%** |
| a) Español | 5031 | 1797 | 25.1% |
| b) Inglés | 565 | 274 | 2.8% |
| c) Otros | 23 | 15 | .1% |
| **LECTURA EN VOZ ALTA** | 464 | 234 | **2.3%** |

La **Lectura en voz alta** consta del 2.3% en el corpus total, cuenta con 464 *tokens* y 234 *types*, a continuación una tabla con los primeros puestos en *types*:

Tabla 3.17. Types *con mayor frecuencia en lectura en voz alta*

| Lugar | Frecuencia de aparición | Palabra |
|---|---|---|
| 1° | 23 | de |
| 2° | 13 | se |
| 3° | 12 | el |
| 4° | 12 | en |
| 5° | 12 | la |
| 6° | 11 | a |
| 7° | 9 | que |
| 8° | 9 | un |
| 9° | 8 | las |
| 10° | 37 | como |

Al igual que en la tabla de **Lectura en silencio**, la tabla de **Lectura en voz alta** sólo contiene dos palabras con significado léxico y el resto son preposiciones y artículos.

La densidad léxica de Lectura en voz alta es de .50, lo que indica que por cada cien palabras la mitad son vocablos distintos. A continuación se presenta una tabla con las palabras con significado léxico y en inglés tanto de **Lectura en voz alta** como de **Lectura en silencio**:

Tabla 3.18. *Palabras más frecuentes con significado léxico en Lectura*

| Lugar | Lugar en Tabla | Frecuencia de aparición | Palabra con contenido léxico y en inglés |
|---|---|---|---|
| 1° | 26° | 22 | *Google* |
| 2° | 29° | 20 | *You tube* |
| 3° | 34° | 16 | gusta |
| 4° | 40° | 14 | buscar |
| 5° | 41° | 14 | *day* |
| 6° | 45° | 14 | web |
| 7° | 47° | 13 | *INXS* |
| 8° | 48° | 13 | opciones |
| 9° | 50° | 13 | *perfect* |
| 10° | 59° | 11 | mundo |

Algunas palabras con contenido léxico o palabras-forma, dentro de la tabla son sustantivos nominales como *Google, You Tube* e *INXS*, todos estos nombres, aunque nombres propios, no violan el código de privacidad, pues son nombres comunes y conocidos, no contienen información privada, en los dos primeros casos

que ocupan el primer y segundo lugar en la tabla, son nombres de buscadores en Internet y el tercero, es el nombre de un grupo musical.

La mitad de las palabras con contenido léxico en la tabla se encuentran en inglés, tres que son nombres propios, un sustantivo y un adjetivo calificativo.

En la sección de **Lectura** encontramos no sólo textos en español o inglés, también encontramos un .4% del total de palabras leídas en otros idiomas. Se diferenció entre la **Lectura en voz alta** y la **Lectura en silencio** pues se consideró que lo que se lee en voz alta es para encontrar una afirmación en lo que se lee, es decir, para comprender cabalmente; en este caso fueron palabras encontradas en las indicaciones de una cámara fotográfica y un texto donde se da información científica además de exclamaciones y quejas. Si tomamos la sección de **Lectura en voz alta** y **Lectura en silencio** como un 100% obtendríamos la siguiente gráfica:

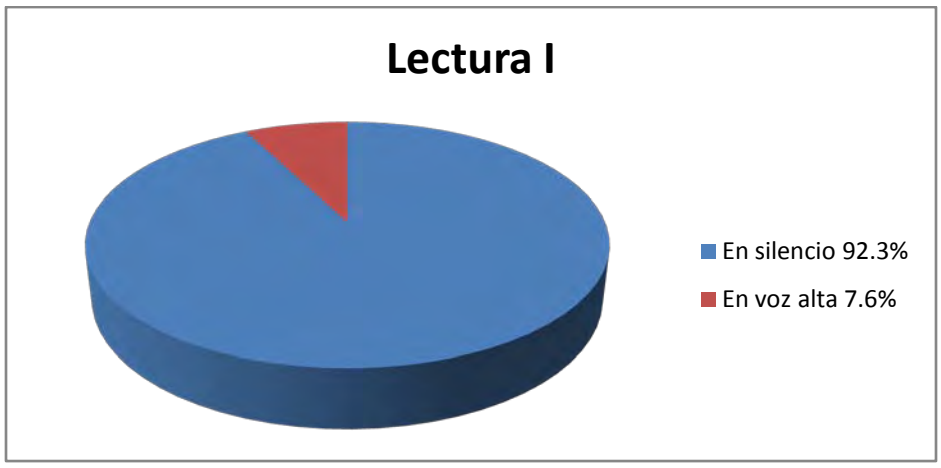



Figura 3.6. *Lectura en silencio* vs. *Lectura en voz alta*

Y si graficáramos los dos tipos de lecturas por separado en comparación con el resto del corpus obtendríamos lo siguiente:

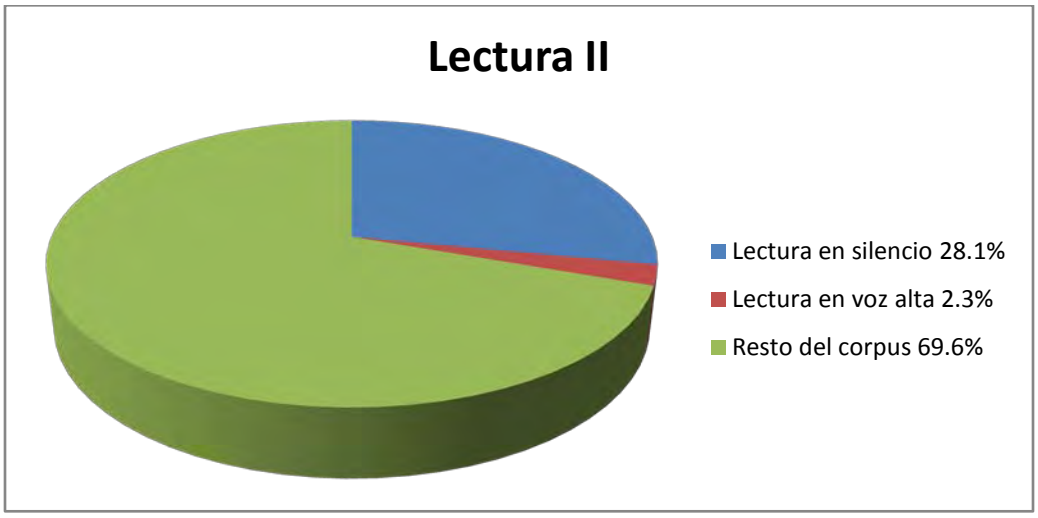



Figura 3.7. *Lectura en silencio, Lectura en voz alta* vs. *Resto del corpus*

Haciendo una comparación entre **Lectura** y **Oralidad** encontramos que es exactamente el mismo porcentaje dentro del total del corpus, ambos obtuvieron un 30.4% pues en el caso de la lectura se juntaron los porcentajes de la que se hizo en silencio 28.1% y la que se hizo en voz alta 2.3%.


Figura 3.8. *Comparación entre Oralidad, Lectura total y Resto del corpus.*

La división de **Lectura en silencio** se hizo con base en los idiomas leídos: Español con 5031 *tokens* y 1797 *types*; Inglés 565 *tokens*, 274 *types;* y Otros con 23 *tokens* y 15 *types*, dentro de esta sección se encuentran palabras como *Sony* y *Kyoto*, básicamente fueron nombres de marcas las que se leyeron en silencio y por separado, es decir, no se encontraban dentro de un texto que les diera contexto. Las palabras que se leyeron en español, no todas son del idioma español, algunas son nombres ingleses como *Gymglish* o *Babylon*; esto se decidió con base en el contexto, pues como estas palabras se encontraron dentro de un texto o contexto en español se tomaron en cuenta como parte del corpus, por ejemplo, la palabra *Youtube* se encontró repetidas veces dentro de la página de *Youtube* pero esta página se leyó en español, aunque poseía palabras en inglés u otra lengua, también se incluyeron

en el corpus de texto leído en español las abreviaturas de nombres propios de canales televisivos en inglés como *Anipl* o *Filmz* por el contexto en donde se leyeron. Dentro de este corpus en inglés se encontraron palabras del libro de Labov *Modelos sociolingüísticos* en donde se habla sobre un tipo de inglés en especial, estas palabras –aunque no inglés estándar- también se incluyeron en esa sección.

Tabla 3.19. *Lectura en división por idiomas*

| **Idiomas** | **Tokens** | **Types** | **% en lectura en silencio** |
|---|---|---|---|
| *Español* | 5031 | 1797 | 89.5% |
| *Inglés* | 565 | 274 | 10% |
| *Otros* | 23 | 15 | .4% |
| *Total* | 5619 | 2043 | 100% |

El porcentaje que se consiguió en esta sección de inglés fue de un diez por ciento, este número no es despreciableen cuanto al input leído, es decir, tiene un peso importante. Como ya se había mencionado este idioma es bastante recurrido en nuestro contexto, ya por la globalización, ya por la cercanía geográfica que tiene México con Estados Unidos. Esta proximidad que se tiene con el inglés provoca una irremediable penetración de este idioma en el nuestro.

## 3.6. MÚSICA

La tabla general está dividida en **Oralidad**, en **Escritura**, **Lectura**, **Híbridos**, **Música y Ruido de fondo**; esta última división no fue incluida en **Oralidad** y fue tomada aparte por considerar que las canciones (aunque obviamente son orales) se encuentran en un ambiente distinto al de la **Oralidad** de los diálogos o monólogos, por no estar las canciones dirigidas a alguien y porque, muchas veces, no se les dedica la atención que se le da a algo articulado y estructurado por nosotros mismos, pues estamos repitiendo palabras que además llevan una tonada, aspecto que fomenta la memorización de esas palabras de manera más sencilla.

La sección **Música** consta de 4748 *tokens* y 782 *types* obteniendo un 23.7% en esta sección del corpus.

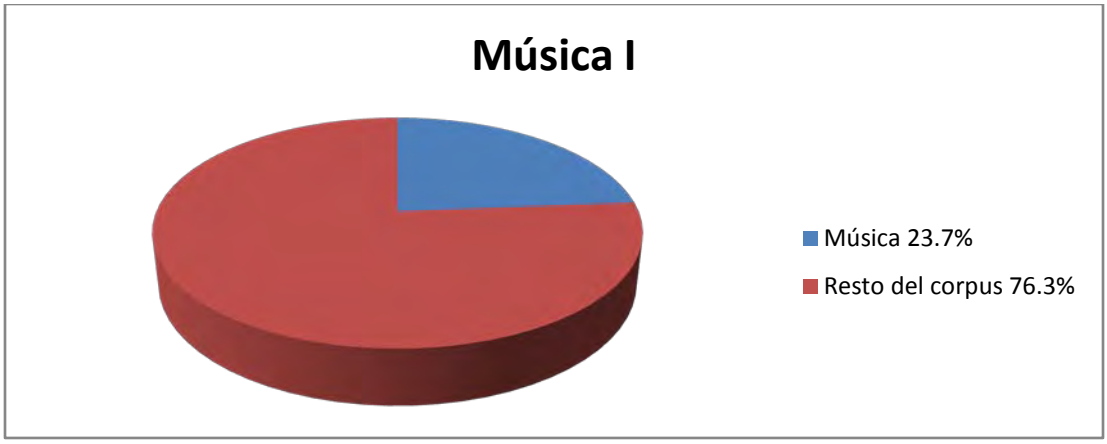



Figura 3.9. *Música* vs. *Resto del corpus*

Esta sección fue dividida en canciones cantadas y escuchadas y a la vez éstas fueron divididas en **Español** e **Inglés**, respecto a las canciones **Cantadas** encontramos un .8% con 172 *tokens* y 82 *types*, dentro de este resultado encontramos 36 *tokens* en español y 136 *tokens* en inglés y dentro de la sección de escuchada encontramos 721*tokens* en español y 3855 *tokens* en inglés, esto demuestra que tanto lo que se escucha como lo que se canta de la radio está en inglés, esto se debe tal vez a la influencia globalizadora que proviene del habla inglesa. Esta influencia evidente del inglés ha traído como consecuencia cambios en el léxico de la población general que podría incrementarse considerablemente en los próximos años por la entrada cultural estadounidense sin filtros al país, podría cambiar por completo el español --al menos eso parece ocurrir con el español chicano (v. al respecto algunos de los trabajos compilados en Lacorte & Lemann 2009).

Esta sección posee un 23.7% del total del corpus, ocupa el tercer lugar después de Oralidad y Lectura en voz baja, la música, aunque no muy importante, se le presta atención lo suficiente como para cambiarla cuando alguna canción no le gusta a la informante-investigadora o para cantarla cuando la canción le gusta. La presencia de la música en este trabajo es sólo un ejemplo de lo que podría ocurrir con un estudiante de universidad promedio.

Si tomamos en cuenta la sección **Música** como un 100%, la cantada constaría con un 3.6% y la escuchada con un 96.3%.

Tabla 3.20. *Música*

| **TIPO TEXTUAL** | **Total Tokens** | **Total Types** | **Porcentaje** |
|---|---|---|---|
| *MÚSICA* | 4748 | 782 | **100%** |
| *Cantada* | 172 | 82 | 3.6% |
| *a) Español* | 36 | 26 | .75% |
| *b) inglés* | 136 | 56 | 2.8% |
| *Escuchada* | 4576 | 785 | 96.3% |
| *a) Español* | 721 | 207 | 15.1% |
| *b) Inglés* | 3855 | 578 | 81.1% |

Separando el total de **Música** en español e inglés obtendríamos una tabla como la siguiente:

Tabla 3.21. *Música en división por idiomas*

| **TIPO TEXTUAL** | **Total Tokens** | **Porcentaje** |
|---|---|---|
| *MÚSICA* | 4748 | **100%** |
| *Español* | 757 | 15.9% |
| *Inglés* | 3991 | 84.1% |

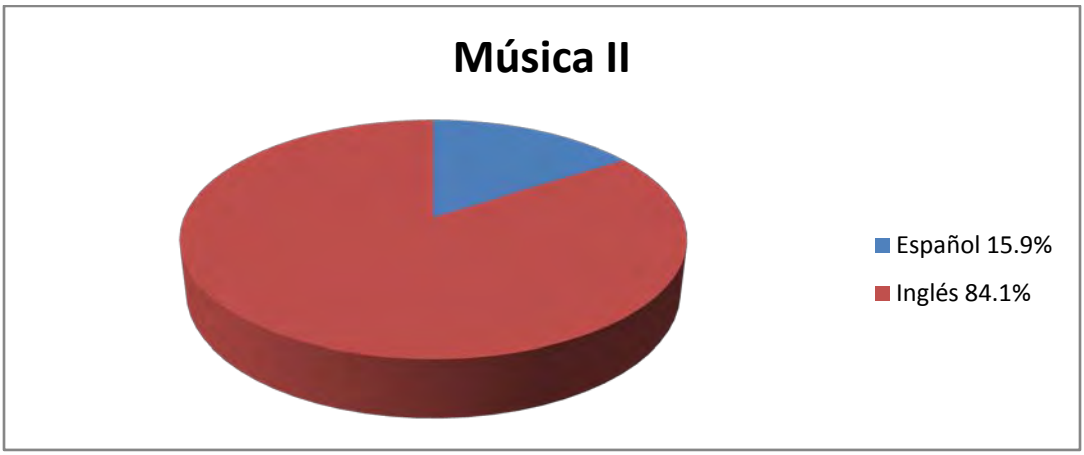



Figura 3.10. *Música en español* vs. *Música en inglés*

## 3.7. RUIDO DE FONDO

A los anuncios de la radio, Internet y conversaciones en las que no se vio involucrada la informante se le llamó **Ruido de Fondo**, a estas palabras no se les prestó atención consciente por no representar nada significativo para la informante. La densidad léxica de esta sección corresponde a .51.

Tabla 3.22. *Ruido de fondo*

| **Tipo Textual** | **Tokens** | **Types** | **Peso % en corpus EN CUANTO A TOKENS** |
|---|---|---|---|
| *Ruido de fondo* | 373 | 192 | 2.1% |
| *Total de corpus* | 19989 | 4178 | 100% |

## 3.8. INPUTY OUTPUT

Dentro del **Input** encontramos todas las palabras que fueron dirigidas al informante, esto es, dentro de la **Oralidad** lo que en el **Diálogo**, las personas que no son la informante  dijeron, lo que expresaron; el **Input** cuenta con un 75.8% dentro del corpus total de *tokens*; este apartado toma en cuenta todo lo que se leyó que fue de un 30.4%, todo lo que en el diálogo se le dijo a la informante que fue de un 14%, también se tomó en cuenta la sección de Híbridos con un 11.8%, la música que escuchó 22.8% y los anuncios de radio y conversaciones lejanas 2.1%. El **Input** suma un total de tres cuartas partes del corpus total, lo que deja al output con un 24.1% dentro del que el 14.3% es lo que habló el informante en el diálogo, el 2% el monólogo, un 1.6% lo que escribió y el .8% lo que cantó, en el **Input** encontramos un total de 3629 *types*. La densidad léxica del total de Input es de .23. En este ejemplo de input se encuentra que un universitario promedio recibe en cuanto a la lengua un 75%, lo que deja sólo un 25% a lo que expone con la lengua.

Tabla 3.23. *Input*

| Tipo textual para INPUT | Total tokens | Total types | Peso % en corpus |
|---|---|---|---|
| *Oralidad/ Diálogo/ Otros* | 2803 | 738 | 14% |
| *Híbridos* | 2377 | 927 | 11.8% |
| *Lectura silencio* | 5619 | 2043 | 28.1% |
| *Lectura en voz alta* | 464 | 234 | 2.3% |
| *Música/ Escuchada* | 4576 | 785 | 22.8% |
| *Ruido de fondo* | 373 | 192 | 2.1% |
| *TOTAL INPUT* | 15158 | 3629 | 75.8% |

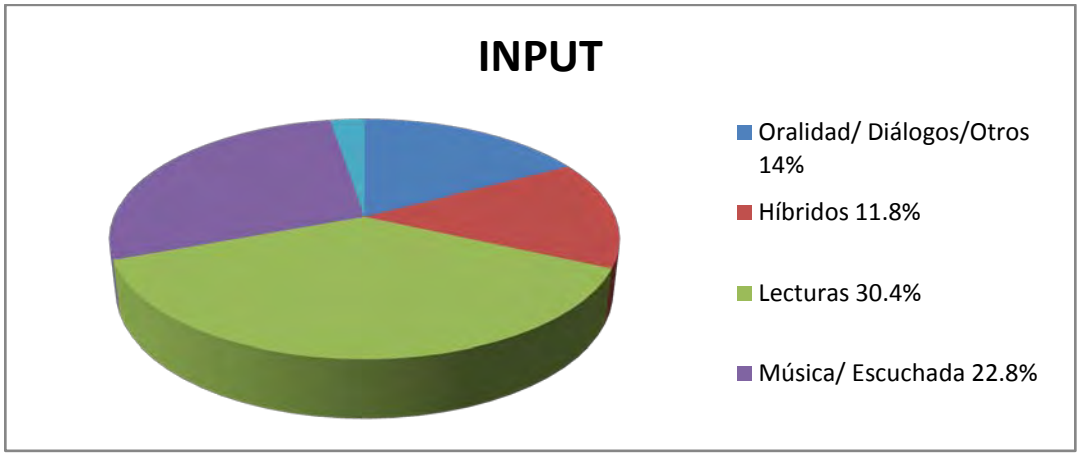



Figura 3.11. *Comparación de los elementos de Input*

Tabla 3.24. Types *con mayor frecuencia en Input*

| **Lugar** | **Frecuencia de aparición** | **Palabra** |
|---|---|---|
| 1° | 441 | de |
| 2° | 350 | a |
| 3° | 334 | que |
| 4° | 266 | la |
| 5° | 240 | no |
| 6° | 229 | y |
| 7° | 208 | el |
| 8° | 207 | *you* |
| 9° | 193 | *the* |
| 10° | 189 | en |

Dentro de estos resultados encontramos que la palabra más recurrente es "de" y que los puestos octavo y noveno están ocupados por palabras en inglés, a continuación la tabla con las palabras más frecuentes con significado léxico dentro del Input.

Tabla 3.25. *Palabras más frecuentes con contenido léxico Input*

| Lugar | Lugar en Tabla | Frecuencia de aparición | Palabra con contenido léxico yen inglés |
|---|---|---|---|
| 1° | 41° | 49 | *red* |
| 2° | 54° | 38 | *time* |
| 3° | 60° | 32 | *celebrate* |
| 4° | 64° | 31 | *free* |
| 5° | 68° | 29 | *feel* |
| 6° | 69° | 29 | *heart* |
| 7° | 77° | 25 | gusta |
| 8° | 79° | 25 | *love* |
| 9° | 83° | 25 | *wine* |
| 10° | 84° | 26 | *disappear* |

El 90% de la tabla son palabras en inglés, esto muestra que además de ser mucho mayor el número de palabras en el Input, las más repetidas con contenido léxico están en inglés.

Ahora se mostrará la tabla con los *types* más frecuentes en la sección de Output.

Tabla 3.26. Types *más frecuentes en Output*

| **Lugar** | **Frecuencia de aparición** | **Palabra** |
|---|---|---|
| 1° | 176 | no |
| 2° | 121 | que |
| 3° | 109 | y |
| 4° | 107 | a |
| 5° | 103 | qué |
| 6° | 100 | de |
| 7° | 93 | es |
| 8° | 80 | el |
| 9° | 66 | la |
| 10° | 57 | me |

Tabla 3.27. *Palabras más frecuentes con contenido léxico Output*

| Lugar | Lugar en Tabla | Frecuencia de aparición | Palabra con contenido léxico |
|---|---|---|---|
| 1° | 36° | 23 | decir |
| 2° | 44° | 17 | saber |
| 3° | 49° | 15 | creer |
| 4° | 57° | 12 | cosas |
| 5° | 62° | 11 | horrible |
| 6° | 78° | 8 | gimnasio |
| 7° | 85° | 7 | *day* |
| 8° | 95° | 7 | taxi |
| 9° | 101° | 6 | amiga |
| 10° | 103° | 6 | cámara |

Tabla 3.28. *Output*

| **Tipo textual para OUTPUT** | **Total tokens** | **Total types** | **Peso % en corpus** |
|---|---|---|---|
| *Oralidad/ Diálogo/ Yael* | 2864 | 704 | 14.3% |
| *Oralidad/ Monólogo* | 419 | 178 | 2% |
| *Escritura* | 322 | 196 | 1.6% |
| *Música/ Cantada* | 172 | 82 | .8% |
| *TOTAL OUTPUT* | 4831 | 1112 | 24.1% |

El Output está constituido por lo que dijo la informante que constó de un 14.3% del corpus total, lo que dijo en Monólogo con 2%, todo lo que escribió con 1.6% y todo lo que cantó con .8%, esto suma un 24.1% del corpus total. Menos de un cuarto de lo que ocurre en un día es emitido por la informante. La densidad léxica es de .23, la misma que se obtuvo en el Input. Con estos resultados se observa que el Output está orientado a la oralidad mientras que el Input, por el contrario, se orienta a la Lectura; lo que se emplea en el Output tienen, básicamente, una naturaleza de audio y el Input de vista, es decir, está orientado al lenguaje escrito.

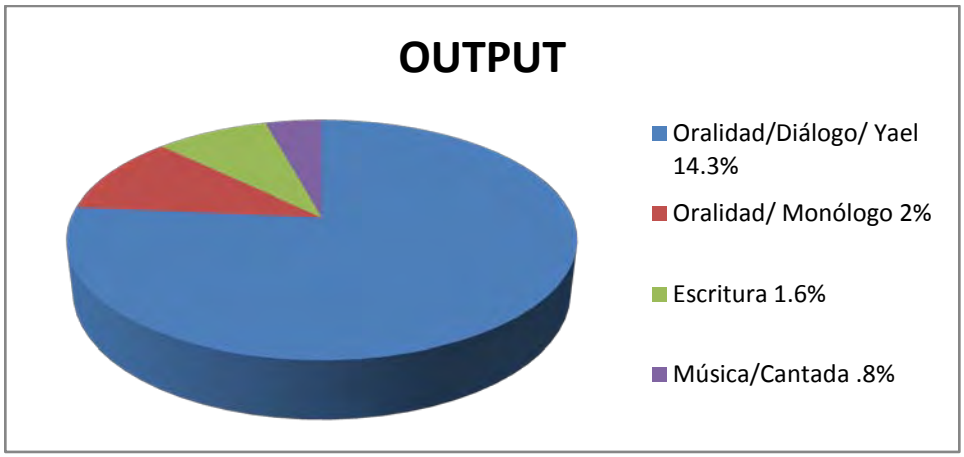



Figura 3.12. *Comparación entre los componentes de Output*

Tabla 3.29. *Input y Output*

| | **Total Tokens** | **Total Types** | **Peso % en corpus EN CUANTO A TOKENS** |
|---|---|---|---|
| *INPUT* | 15158 | 3629 | 75.8% |
| *OUTPUT* | 4831 | 1112 | 24.1% |
| *Total de corpus* | 19989 | 4178 | 100% |

Estos resultados muestran que la informante recibe información tres veces más de la que expresa;si este fuera el caso con el resto de la población estudiantil universitaria, indicaría que somos más receptores que emisores, recibimos más información para procesar en nuestra mente de lo que expresa nuestra mente misma, aunque no estamos contando todo aquello que se dice sin hablar, escribir o leer, ni todo aquello que pensamos.

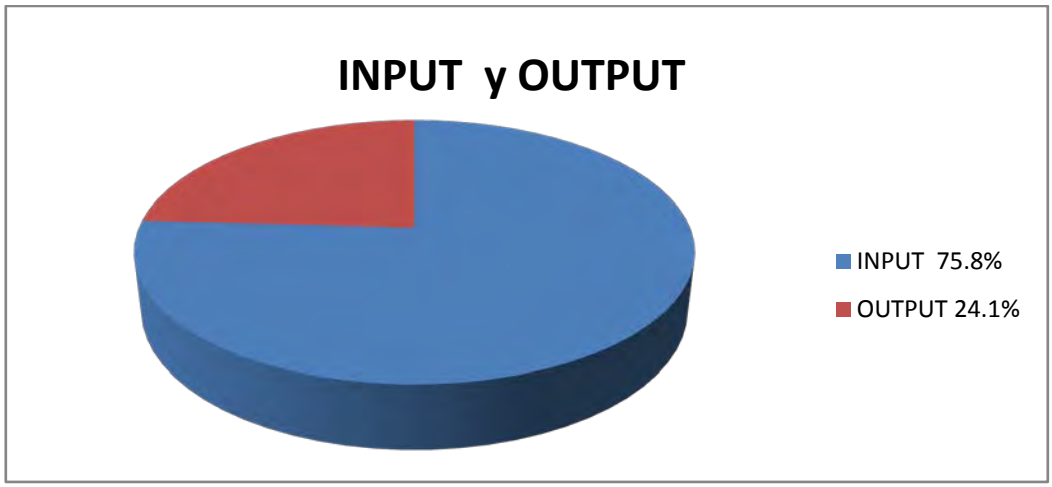



Figura 3.13. *Input* vs. *Output*

## 3.9. RESULTADOS GENERALES

Tabla 3.30. *Resultados generales según tipo textual*

| TIPO TEXTUAL | TOTAL TOKENS | TOTAL TYPES | PESO % EN CORPUS (EN CUANTO A TOKENS) |
|---|---|---|---|
| *ORALIDAD* | 6086 | 1162 | **30.4%** |
| *Diálogo* | 5667 | 1099 | 28.3% |
| *a) Yael* | 2864 | 704 | 14.3% |
| *b) Otros* | 2803 | 738 | 14% |
| *Monólogo* | 419 | 178 | 2% |
| *ESCRITURA* | 322 | 196 | **1.6%** |
| *LECTURA SILENCIO* | 5619 | 2043 | **28.1%** |
| *a) Español* | 5031 | 1797 | 25.1% |
| *b) Inglés* | 565 | 274 | 2.8% |
| *c) Otros* | 23 | 15 | .1% |
| *LECTURA EN VOZ ALTA* | 464 | 234 | **2.3%** |
| *MÚSICA* | 4748 | 782 | **23.7%** |
| *Cantada* | 172 | 82 | .8% |
| *c) Español* | 36 | 26 | .18% |
| *d) inglés* | 136 | 56 | .68% |
| *Escuchada* | 4576 | 785 | 22.8% |
| *c) Español* | 721 | 207 | 3.6% |
| *d) Inglés* | 3855 | 578 | 19.2% |
| *HÍBRIDOS* | 2377 | 927 | **11.8%** |
| *RUIDO DE FONDO* | 373 | 192 | **2.1%** |
| *INPUT TOTAL* | 15158 | 3629 | 75.8% |
| *OUTPUT TOTAL* | 4831 | 1112 | 24.1% |
| *GRAN TOTAL* | 19989 | 4178 | 100% |

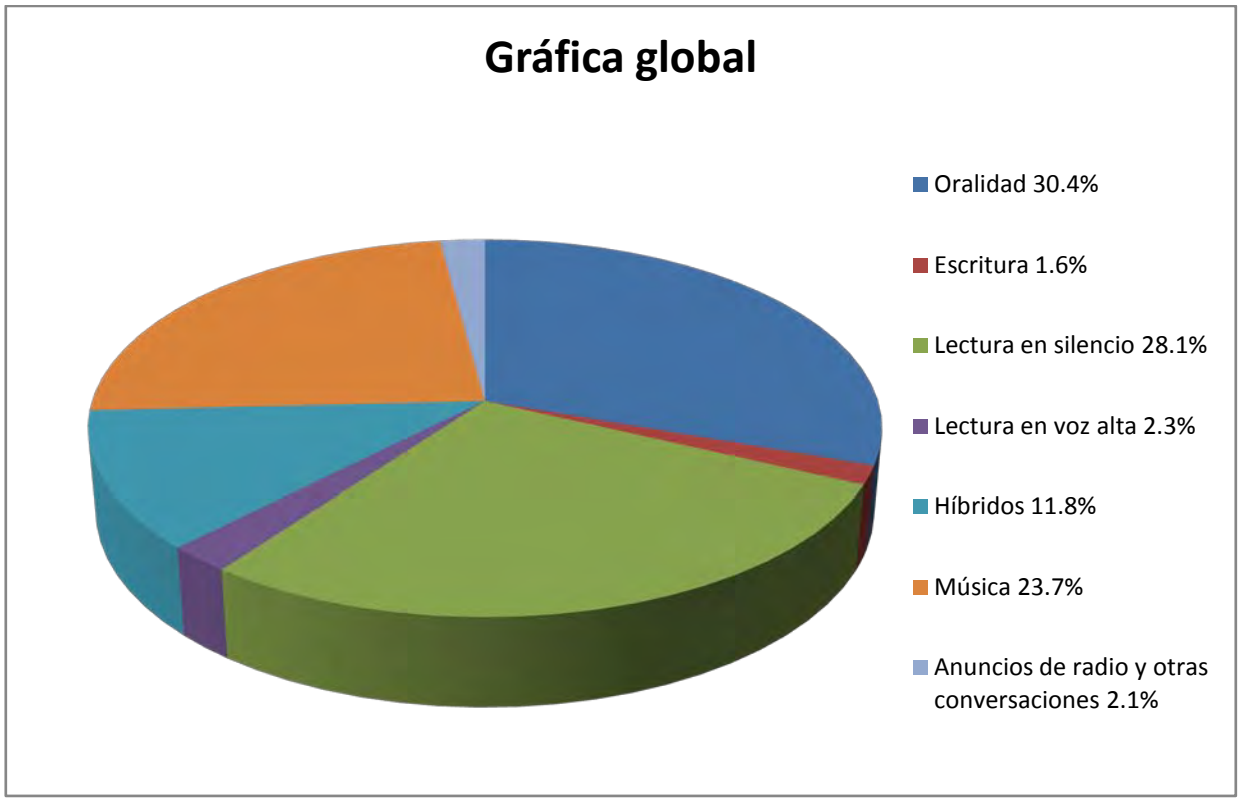



Figura 3.14. *Gráfica global*

Tabla 3.31. Types *más frecuentes en Total*

| Lugar | Frecuencia de aparición | Palabra |
|---|---|---|
| 1° | 559 | de |
| 2° | 483 | que |
| 3° | 469 | a |
| 4° | 431 | no |
| 5° | 362 | y |
| 6° | 339 | la |
| 7° | 303 | el |
| 8° | 249 | es |
| 9° | 248 | en |
| 10° | 214 | *you* |

En general, estos resultados coinciden con la lista de español general del CREA (RAE, 2012). También existe coincidencia respecto al español mexicano, según se expone en el libro de Luis Fernando Lara (1979), en la tabla con las 100 palabras más recurridas del listado del Corpus del Español Mexicano Contemporáneo (CEMC). Finalmente, en el libro *Variación social del léxico disponible en la ciudad de Málaga* (Ávila Muñoz & Villena Ponsoda, 2010) se muestra una tabla de los 100 vocablos más usados en el estudio sobre el léxico utilizado en dicha ciudad española. A continuación se presentan las primeros diez palabras de las tres tablas mencionadas:

Tabla 3.32. Tokens *de mayor frecuencia en el CREA*

| Lugar | Palabra |
|---|---|
| 1° | de |
| 2° | la |
| 3° | que |
| 4° | el |
| 5° | en |
| 6° | y |
| 7° | a |
| 8° | los |
| 9° | se |
| 10° | del |

Tabla 3.33. Tokens *con mayor frecuencia en CEMC*

| Lugar | Palabra |
|---|---|
| 1° | la |
| 2° | el |
| 3° | de |
| 4° | y |
| 5° | que |
| 6° | en |
| 7° | a |
| 8° | se |
| 9° | no |
| 10° | ser |

Tabla3.34.Types*de mayor uso en estudio de Málaga (oralidad)*

| Lugar | Palabra |
|---|---|
| 1° | sí |
| 2° | con |
| 3° | pero |
| 4° | pues |
| 5° | ese |
| 6° | ya |
| 7° | hacer |
| 8° | por |
| 9° | el |
| 10° | decir |

Estas tres tablas, unta con la generada en el presente estudio, en su mayor parte están compuestas por palabras sin contenido léxico, a excepción del décimo lugar en la lista de Málaga, que es *decir*[8].

Entre la tabla del presente trabajo y la del CREA hay un 70% de coincidencia en cuanto a las palabras y concuerdan en la primera posición de la tabla con la palabra *de.*

Entre la tabla de CEMC y la de este trabajo se comparten ocho de las 10 palabras[9] mientras que con la de Málaga sólo una palabra, el artículo "el". Con los datos que corresponden al habla general del español, es decir, el CREA, encontramos grandes parecidos con la de este trabajo y con la del CEMC, en la tabla de Málaga hay discrepancias pero en líneas generales se comprueba que los vocablos más recurridos por una persona carecen de contenido léxico. El parecido de la tabla de este trabajo y las demás, comprueban que los resultados que se obtuvieron en este trabajo son una herramienta útil por contener datos fehacientes.

A continuación se presentan cuatro tablas con las palabras más frecuentes con contenido léxico en el CREA (RAE, 2012), en Lara (1979), en Ávila Muñoz & Villena Ponsoda (2010) y en este trabajo.

[8] No se toman en cuenta los verbos copulativos.

[9] Podría ser el 90% si se toma exactamente como la misma palabra el verbo "ser" y el verbo "es".

Tabla 3.35. *Palabras de contenido léxico con mayor uso en el CREA*

| Lugar | Palabra |
|---|---|
| 1° | había |
| 2° | años |
| 3° | puede |
| 4° | parte |
| 5° | tiene |
| 6° | bien |
| 7° | tiempo |
| 8° | vida |
| 9° | hace |
| 10° | gobierno |

Tabla 3.36. *Palabras de contenido léxico con mayor frecuencia en CMEC*

| Lugar | Palabra |
|---|---|
| 1° | tener |
| 2° | hacer |
| 3° | decir |
| 4° | poder |
| 5° | ir |
| 6° | ver |
| 7° | dar |
| 8° | año |
| 9° | saber |
| 10° | querer |

Tabla 3.37. *Palabras de contenido léxico con mayor uso en estudio de Málaga*

| Lugar | Palabra |
|---|---|
| 1° | hacer |
| 2° | decir |
| 3° | bueno |
| 4° | ver |
| 5° | saber |
| 6° | año |
| 7° | venir |
| 8° | poder |
| 9° | cosa |
| 10° | pasar |

Tabla 3.38. *Palabras de contenido léxico más frecuentes en el presente corpus*

| Lugar | Lugar en Tabla | Frecuencia de aparición | Palabra con contenido léxico |
|---|---|---|---|
| 1° | 36° | 73 | decir |
| 2° | 64° | 45 | saber |
| 3° | 70° | 42 | *time* |
| 4° | 77° | 35 | *day* |
| 5° | 87° | 32 | *celebrate* |
| 6° | 88° | 32 | *free* |
| 7° | 91° | 30 | *perfect* |
| 8° | 96° | 29 | *feel* |
| 9° | 97° | 29 | gustar |
| 10° | 98° | 29 | *heart* |

En las primeras tres tablas se incluyeron verbos como *hacer* y *poder* por falta del corpus de estos estudios y por ende del contexto de cada verbo. En las cuatrotablas

encontramos sustantivos que se refieren al tiempo y los verbos *decir* y *saber* y en cuanto a la tabla de este trabajo se encuentra en inglés el 70%, lo que indica que hay un peso indudable del inglés en la interacción que tiene la informante-investigadora con la lengua inglesa. Este fenómeno muy probablemente llegue a afectar más el léxico de las futuras generaciones en nuestro país, dada la enorme influencia del inglés sobre todo a partir de los fuertes procesos de globalización económica y cultural que estamos viviendo.

### 3.10. CONCLUSIONES GENERALES

La tabla de contenido léxico indica las diez palabras-forma que más se repitieron del total del corpus, dentro de estas se encuentran siete en inglés, esto se debe a las repeticiones que las canciones contienen, por esto mismo la tabla de palabras con contenido léxico del Input contiene un 90% de contenido en inglés.

Son recurrentes las palabras *dice* y *dijo* dentro de las secciones Híbrido, Oralidad y Escritura, por esto es que *dijo* es una de las tres palabras en español con contenido léxico del corpus total. Además de esta palabra las palabras que más se repiten en Oralidad, Escritura, Híbrido son los verbos: *saber*, *pensar*, *creer*, los tres verbos son procesos mentales, no conllevan acciones físicas, son meramente intelectuales.

Se encontraron tipos textuales que no era posible categorizar en los ya conocidos, e incluso hubo algunos que cuentan con características un tanto distintas del grupo de tipo textual en donde se incluyeron, por ejemplo: el monólogo, la lectura en voz alta y un pseudo diálogo, este último ni siquiera se separó en la tabla general del monólogo por no representar un porcentaje significativo y por la dificultad para seleccionarlo, éste tipo textual se descubrió al concientizar la forma de hablar dentro del mismo monólogo, se consideró al pseudo diálogo como el decir palabras que indirectamente están dirigidas a un receptor pero que se encuentran dentro del tono y dentro del monólogo mismo, de este pseudo diálogo se espera una posible respuesta, pero si no se recibe, no se le da importancia.

En los resultados de Input y Output se encontró una clara diferencia a favor del Input pues éste contó con un 75.8% del total del corpus, este dato podría hacer que se realice la hipótesis de que todos los informantes reciben más de la lengua de lo que proporcionan a partir de ella, esto se debe a que todo el tiempo estamos escuchando, incluso lo que no se desea; por el contrario, cuando se necesita ya sea escribir o hablar, se escoge un momento preciso y las palabras para expresar con precisión lo que se desea.

La relación que se encontró entre hablantes en el Diálogo es, como ya se mencionó anteriormente, equitativa, en esta interacción hay traslapes, pero no son impedimento para que haya atención por parte de cada participante en cuanto a lo que el otro expresa; en el inicio se intuía por observación, antes de hacer la recopilación y el análisis del diálogo, que este porcentaje sería equitativo, como efectivamente se demostró tras el análisis cuantitativo con *Ant Conc*.

A diferencia de lo que se creía antes de hacer la recopilación de datos en el sentido de que el corpus de Escritura sería extenso, los datos mostraron un corpus bajo, sólo el 1.6% dentro del total, fue incluso más alto que la sección denominada Ruido de fondo, por ser escaso lo que se escribió pero fue el más denso en su léxico con un .60. En cuanto a densidad la siguiente sección con mayor en densidad léxica fue Ruido de fondo con .51.

Con respecto a la densidad léxica, total del corpus del que se obtuvo un .20, este resultado refleja que las secciones con mayor densidad fueron pobres en comparación con las de poca densidad como Música.

Si se hace una comparación entre Oralidad y Escritura se obtiene una gráfica como esta:

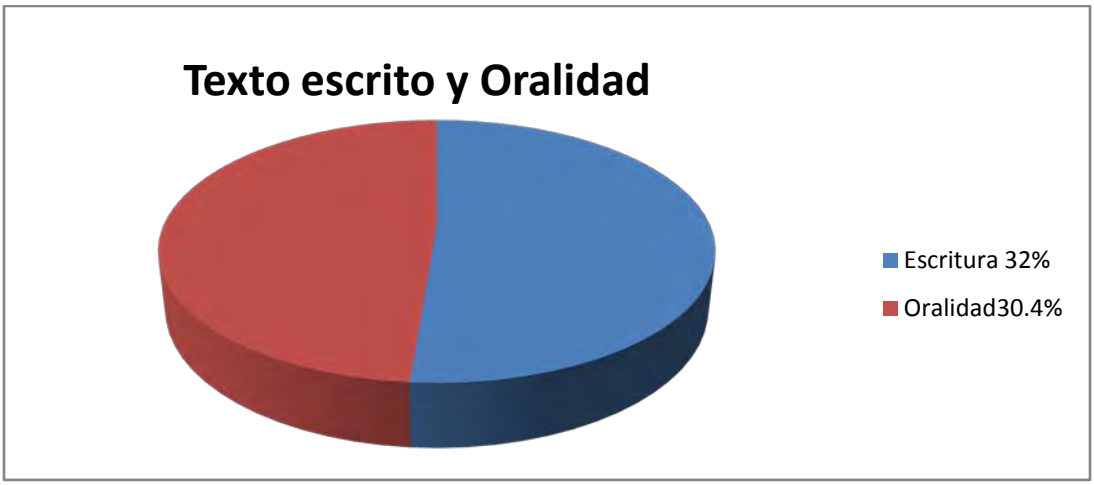



Figura 3.15 *Texto escrito* vs. *Oralidad*

Esta comparación es de suma importancia pues revela el peso real que le damos a los textos leídos en comparación con el que le damos a la oralidad en un día común. Aunque este resultado no sea el de una persona promedio en la ciudad de México, sí podría ser representativa del sector universitario promedio. Este resultado podría hacer que se reflexione más acerca de la interacción que un individuo común tiene con la lengua y, siguiendo estos parámetros,se creen corpora novedosos para mantener un registro y seguimiento más fidedigno de lo que ocurre con la lengua. Los datos recogidos por la RAE para los corpora que son autoridad en el habla española, constan de un 90% de textos escritos, mientras que sólo un 10% son datos recogidos de registros orales. Comparando los resultados de este ensayo socialmente realista y lo que se toma en cuenta para el buen uso del español encontramos una gran diferencia pues en este ensayo hay una casi

igualdad entre ambas partes: la diferencia entre estas es de menos del 2%, creando así una propuesta que es la de que en vez de darle sólo el 10% de importancia a la oralidad se le dé un 30% y que a los textos escritos se les dé también un 30% pues es lo que los resultados señalan dejando al resto entre Música, Híbridos y Ruido de fondo.

# CONCLUSIONES Y REFLEXIONES FINALES

La sociolingüística busca obtener el mejor dato de la lengua, esto es, obtener la lengua vernácula, pues a través de estos datos pueden ser analizadas a detalle las características de la lengua que realmente utilizamos. Es complicado registrar la lengua vernácula debido a la *Paradoja del observador.* Ésta dicta que, por un lado, se quiere estudiar el lenguaje en su forma más natural, cuando no hay observador (porque el observador sesga la interacción lingüística); sin embargo, no puede haber observación del dato lingüístico sin observador, y por lo tanto, el sesgo es siempre inevitable. Un efecto de dicha paradoja es que mientras mejores sean las condiciones para el registro de la lengua (por ejemplo, en un laboratorio de fonética, utilizando técnicas experimentales), más nos alejamos de la lengua natural y mientras menores sean las condiciones para obtener un buen audio (por ejemplo, utilizando una grabadora oculta), el informante será más natural en cuanto a su lengua. Este problema puede ser relativamente superado al utilizar como método la grabación secreta, como se hizo en este trabajo, en el que la investigadora-informante se grabó a sí misma y a todos aquellos que interactuaron con ella durante un día completo. Aunque la investigadora-informante pudo haber cambiado algo de su forma de hablar, los informantes que no supieron de la grabación no hicieron ningún cambio en su

lengua natural consiguiendo así superar la paradoja del observador. Por tratarse de un auto-registro se presentaron menos problemas de los que se pueden tener registrando a otra persona, se tuvo más control de la grabadora y de lo que se leía o lo que se podría hacer con otro informante. Los datos que se obtuvieron en un día revelaron porcentajes, densidades léxicas y tipos textuales que permitieron un acercamiento analítico a la lengua natural que tiene distintos comportamientos dependiendo de la situación en la que el informante se encuentre, a esto William Labov le llama "variantes estilísticas" (Labov 1972).

Se encontró que además de la lectura, habla y escritura, existen formas de recibir y expresar la lengua distintas a las que no se les puede designar completamente alguno de estas asignaturas como, por ejemplo, el tipo textual llamado en este trabajo Híbrido[1], que es una mezcla entre lectura y oralidad pues se escucha al mismo tiempo que se lee. Además de este tipo de expresión, se encontró el Monólogo[2] y el pseudo diálogo[3]. El haber captado y mencionado la existencia de estos tipos textuales en un día común hace que este trabajo se acerque a lo que realmente se hace con la lengua cotidianamente aunque se hayan encontrado relativamente pocos ejemplos.

[1] Este tipo textual contó con el 11.8% del total del corpus.

[2] Que se incluyó en Oralidad y que cuenta con el 2% del total.

[3] Que se incluyó en Monólogo sin hacer diferencias con éste.

Los datos del corpus en cuanto a la Oralidad, indicaron que la participación en los diálogos entablados entre la investigadora y los informantes fue de un casi 50/50, esto indica que la cantidad de palabras que expresan dos personas en una conversación es casi la misma, además de que hay un respeto de turnos entre participantes. Muestran los resultados que hay moderación por parte de ambas partes de una plática, y que hay un límite --tal vez inconsciente-- que se imponen los participantes en una conversación cualquiera en cuanto a la cantidad de palabras pronunciadas. También se encontró que las palabras más frecuentes son casi las mismas para la investigadora-informante y sus interlocutores. Además del diálogo se registró el monólogo que, por observación, la gente casi no practica y mucho menos si se sabe grabada como la investigadora-informante. La Oralidad en total obtuvo un 30.4% del total del corpus al igual que el total de la Lectura cuyas palabras con significado léxico o palabras plenas con mayor frecuencia fueron nombres propios en inglés. La Escritura contó con sólo el 1.6% del total del corpus pero su densidad léxica fue la mayor con un .60, esto indica que las palabras casi no se repitieron, se supone esto porque no hubo oportunidad pues fueron pocas palabras. La sección de Música que contó con un 23.7% del que el 84.1% son palabras en inglés y el resto en español, este tipo de texto cuenta con el nivel de densidad más bajo de .16. Estos datos son la explicación a la baja densidad en el corpus total y a que la

mayor parte de las palabras plenas con mayor frecuencia dentro del total sea en inglés. El total de corpus fue dividido en dos grandes secciones que engloban lo que expresó -output- y lo que recibió –input- la investigadora-informante, respectivamente contienen el 24.1% y el 75.8% lo que indica que es lo que se recibe es tres veces más de lo que se expresa dentro de la lengua. Todos estos datos fueron obtenidos de un corpus realista que no hizo a un lado ningún caso en donde la lengua estuviera involucrada.

Una de las cuestiones en la que se interesa la sociolingüística es el cómo se puede hacer un corpus realista. En general puede decirse que existe cierta falta de realismo social en los *corpora* de la Real Academia de la Lengua Española (RAE). Nos referimos específicamente al Corpus de Referencia del Español Actual (http://corpus.rae.es/creanet.html) y al Corpus Diacrónico del Español (http://corpus.rae.es/cordenet.html) mejor conocidos por sus siglas CREA y CORDE respectivamente y que son ampliamente utilizados en la lingüística hispánica contemporánea. Estos *corpora* toman en cuenta en un 50% a todo el continente americano dejando el restante 50% a la Península Ibérica. Si se toma en cuenta que la cantidad de hablantes sólo en México es del 25% del total de los hablantes de español en el mundo –esto es, casi 100 millones-- (López Morales 2006-2007) mientras que España posee el 10%, ¿por qué no construir un corpus representativo y fiel a estas proporciones demo-linguísticas? Por supuesto, y

como se comenta en Serrano (2011), no es tarea de la RAE hacer este trabajo, pero vale la pena reflexionar sobre un corpus en ese sentido más realista.

Por otra parte, los datos recogidos por la RAE para estos *corpora* constan de un 90% de textos escritos mientras que sólo un 10% son de datos recogidos de registros orales –por cierto, de muy dudosa naturalidad, ya que para México se trata de discursos en la Cámara de Diputados y Senadores. Pero ¿cuánto tiempo pasa un hablante de español interactuando con la lectoescritura y cuánto tiempo pasa interactuando oralmente? A esta pregunta el presente trabajo da una respuesta que, aunque no totalmente generalizable, proporciona un dato real: estos datos muestran que hay una relación bastante equilibrada entre lectoescritura y oralidad, ya que ambos tipos de texto obtuvieron un 30.4% del total del corpus.

Además, el método utilizado proporcionó otros datos de interés en cuanto a los diferentes tipos de interacción lingüística, que no se limitan a lengua hablada *versus* lengua escrita. Para diseñar y elaborar un corpus que muestre la lengua tal como es, es necesario tomar en cuenta lo que ocurre en las distintas comunidades de habla de manera cotidiana, pues ¿con qué criterio se dictamina lo que es la lengua "correcta"? A esta interrogante Luis Fernando Lara en su *Teoría del diccionario monolingüe* (1997) responde*:*

> Los diccionarios del siglo XVII hasta el *Diccionario de Autoridades* de la Academia Española eran… más catálogos simbólicos, representativos, de la calidad del vocabulario literario, restringido por la idea de la lengua imperante, que verdaderas obras de consulta generales. Su simbolismo se dirigía a la legitimación de las lenguas literarias europeas; representaba la lengua como celebración del Estado ante los miembros de la sociedad que participaban en él: la nobleza, los letrados… El resto de la comunidad lingüística quedaba fuera del círculo simbólico en que se elaboraban los diccionarios y tenían sentido: en el mejor de los casos, era un espectador (p. 46).

Esto es, la "autoridad" nunca se obtenía de los hablantes comunes, y con esto es un grave desacierto, sobre todo en el siglo XXI, cuando la lingüística moderna confiere importancia a estos hablantes "comunes".

Al respecto, Lara (1997: 81) cita a Sledd & Ebbitt (1962: 83): "los compiladores dirían sin duda que la autoridad del diccionario se basa en el uso, pero se plantea la cuestión insistentemente: ¿El uso de quién?". Lo que se propone aquí, es que son precisamente los hablantes comunes quienes deberían proporcionar los datos y la "norma" (en el sentido de Coseriu [1958] de "normalidad estadística") para redactar una gramática descriptiva y dejar a las "autoridades" (narradores, poetas, ensayistas y expertos en filología) la

redacción de una gramática normativa[4].

Por supuesto, la constitución de un corpus basado en grabaciones secretas y el seguimiento de una solo hablante durante todo un día se percibe logísticamente complicado, pero al menos proporcionaría una visión realista de lo que ocurre con el lenguaje cotidiano, esto es, una sólida base de datos *lingüísticos primarios* (Himmelmann, 2007) con grandes ventajas para el análisis sistémico y de los procesos de variación sociolingüística y cambios en la estructura de la lengua.

[4] Por supuesto, se debe reconocer que la *Gramática descriptiva de la lengua española* coordinada por Ignacio Bosque y Violeta De Monte (1999) constituyó un gran esfuerzo por redactar una gramática no purista, aunque esté muy alejada todavía de procedimientos empíricos para la obtención del dato, como los que aquí se sugieren (v. la discusión en Serrano 2011: 210-213).

BIBLIOGRAFÍA

Ávila Muñoz, Antonio M. & Juan A. Villena Ponsoda (eds.) (2010). *Variación social del léxico disponible en la ciudad de Málaga. Diccionario y Análisis.* España: Sarriá.

Ávila Muñoz, Antonio M. (2010). "Análisis estadístico comparativo y estudio cualitativo", en Ávila Muñoz& Villena Ponsoda (eds.) (2010:111-176).

Ávila, Raúl (1999). *Estudios de semántica social.* México: El Colegio de México.

Báez Pinal, Gloria Estela, Elizabeth Luna Traill & Alejandra Vigueras Ávila (2005). *Diccionario básico de lingüística*. México: Universidad Nacional Autónoma de México.

Bedmar, Ma. Jesús (1995). "La posesión de la lengua (de E. Coseriu a G. Salvador)", *Revista Española de Lingüística* 25 (1), 87-97.

Berruto Torino, Gaetano (2010). "Structure and dynamics of a language space" en *Language and Space Volume 1: Theories and Methods.*Editado por Peter Auer y Jürgen Erich Schmidt. Berlin: De Gruyter Mouton; pp. 44-56.

Bosque Ignacio, Demonte, Violeta (1999). *Gramática descriptiva de la lengua española.*Madrid: Espasa Calpe

Chambers, Jack K. & Peter J. Trudgill (1994). *La dialectología*. Barcelona: Visor libros.

Chambers, Jack K. (1995).*Sociolinguistic Theory.*Oxford: Blackwell.

Coseriu, Eugenio (1958). *Sincronía, diacronía e historia*. Madrid: Gredos.

Fishman, Joshua (1974).“Conservación y desplazamiento del idioma como campo de investigación”, en Garvin & Lastra (1974: 375-423).

Garvin, Paul & Yolanda Lastra (eds.) (1974). *Estudios de etnolingüística y sociolingüística.* México: Universidad Nacional Autónoma de México.

Ham, Roberto Chande, Luis Fernando Lara y Ma. Isabel García Hidalgo (1979). *Investigaciones lingüísticas en lexicografía.* México: El Colegio de México.

Haviland, John B. & J. Antonio Flores Farfán (coords.) (2007). *Bases de la documentación lingüística*. México: Instituto Nacional de Lenguas Indígenas.

Himmelmann, Nikolaus P. (2007). “Cap. 1. La documentación lingüística: ¿qué es y para qué sirve?” en Haviland & Flores Farfán (2007: 15-47).

Hymes, Dell (1994). “Hacia etnografías de la comunicación”, en Garvin & Lastra (1974:48-89).

Labov, William (1983). *Modelos sociolingüísticos.* Madrid: Cátedra.

Lacorte, Manuel& Jennifer Leeman (eds.) (2009). *Español en Estados Unidos y otros contextos de contacto. Sociolingüística, ideología y pedagogía/Spanish in the United States and Other Contact Environments.Sociolinguistics, Ideology, and Pedagogy*. Madrid: Vervuert/Iberoamericana.

Lara, Luis Fernando (1997). *Teoría del diccionario monolingüe*. México: El Colegio de México.

Lastra Yolanda & Butragueño Martín (2000). "Modo de vida como factor sociolingüístico en la Ciudad de México", en *Estructuras en contexto. Estudios de variación lingüística.* Edición de Pedro Martín Butragueño México: Colegio de México; pp. 13-39.

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

Moreno Fernández, Francisco (1990).*Metodología sociolingüística*, Madrid: Gredos.

Moreno Fernández, Francisco(1998).*Principios de sociolingüística y sociología del lenguaje,* España: Ariel, 4ta edición, 2009.

Moreno Fernández, Francisco(2011). "Prólogo. Historia, sociedad y lengua", en *Historia sociolingüística de México, vol. I: México prehispánico y colonial.* Rebeca Barriga Villanueva & Pedro Martín Butragueño (dirs.). México: El Colegio de México; pp. 27-39.

Serrano, Julio César (2009)."Procesos de digitalización y transcripción del habla popular de la ciudad de México". Ponencia presentada en *Jornadas Filológicas 2008-2009*. Instituto de Investigaciones Filológicas, UNAM, 9-13 de marzo.

Serrano, Julio César (2011). "Retracción e innovación léxica en español de la ciudad de México 1970-2000",en *Realismo en el análisis de corpus orales*. *Primer coloquio de cambio y variación lingüística.* Pedro Martín Butragueño (ed.). México: El Colegio de México; pp. 191-215.

Silva Corvalán, Carmen (2001). *Sociolingüística y pragmática del español.* Washington, D. C.: Georgetown University Press.

Trudgill, Peter (2000). "Sociolinguística y sociolingüística", en *Estudios de sociolingüística*. Yolanda Lastra (comp.). México: Instituto de Investigaciones Antropológicas, UNAM; pp. 1-37.

Tognini-Bonelli, Elena (2001). *Corpus Linguistics at Work.* Amsterdam-Philadelphia: John Benjamins.

Ullmann, Steven (1972).*Semántica. Introducción a la ciencia del significado.* Madrid: Aguilar.

Schilling-Estes, Natalie (2007)."Sociolinguistic fieldwork", en *Sociolinguistic Variation.Theories, Methods, and Applications.* Robert Bayley & Ceil Lucas (eds.). Nueva York: Cambridge University Press.

Weinreich, Uriel, William Labov y Marvin I. Herzog (1968). "Empirical foundations for a theory of language change", en *Directions for Historical Linguistics.A Symposium*. W. P. Lehmann & Y. Malkiel (eds.). Austin: University of Texas Press, pp. 95-195.

PÁGINAS DE INTERNET CONSULTADAS Y PROGRAMAS DE CÓMPUTO

Anthony, Laurence (2010). *AntConc (version 3.1.4 w).A Freeware Concordance Program for Windows, Macintosh OS X, and Linux*.Waseda University, Japón. Descargable en: http://www.antlab.sci.waseda.ac.jp/software.html.

*Corpus de Referenciadel Español Actual.* Madrid: Real Academia Española de la Lengua. http://corpus.rae.es/creanet.html.Consultada en marzo de 2011.Lista de frecuencias consultada en mayo de 2012 en: http://corpus.rae.es/frec/1000_formas.TXT.

*Corpus Diacrónico del Español.* Madrid: Real Academia Española de la Lengua. http://corpus.rae.es/cordenet.html. Consultada en marzo de 2011.

*LAMAC: Latin American Multichannel Advertising Council.* http://www.lamac.org/mexico/metricas/total-por-tv-paga.Consultada el 23 de agosto de 2011.

# APÉNDICE I

## CRITERIOS DE TRANSCRIPCIÓN[1]

*Unidades*

(a) Turno: X: Tab (sin espacio blanco; el turno empieza siempre con minúscula, y al final no se pone nada en espacial; los turnos se numerarán en la versión definitiva).

(b) Traslapes: [ ] (sin espacio).

(c) Palabra trunca: t- (sin espacio).

*Prosodia*

(a) Énfasis fuerte: ¡¡ !! (sin espacio)

(b) Énfasis moderado: ¡ ! (ídem)

(c) Interrogación: ¿ ? (ídem)

(d) Suspensión voluntaria: … (sin espacio en la palabra a que se adjuntan).

(e) Alargamiento: a: t:: (sin espacio)

(f) Breve: t/ t (sin espacio a la izquierda, un espacio a la derecha de la barra).

[1] Tomados de una versión preliminar para el *Corpus Sociolingüístico de la ciudad de México* (Martín Butragueño y Lastra 2011: 46-62).

(g) Media: // (ídem).

(h) Larga: // (ídem; equivale a un silencio).

Al final de turno no se marca pausa, salvo que exista una buena razón para marcarla.

*Ruidos*

(a) Ruidos: (risa) (clic) (ruido) (carraspeo), etc.

*Fonética*

(a) Ortografía ordinaria

(b) Pronunciación: pues <~pos> (un espacio).

*Observaciones*

(a) Comentarios del analista: ( ) (cuando aparezcan en el texto; comentarios más detallados preferiblemente que hayan en notas). Ejemplos de comentarios: (interrupción de la grabación) (risas de todos) (vacilación) (corrección) (silencio)- para intervalos de 1.2 a 2 segundos- (lapso)- para intervalos de más de 2 segundos-.

(b) Fragmento ininteligible: <…> (sin espacios dentro).

(c) Texto inseguro: <voy a estar allá> (sin espacio).

*Notaciones especiales*

(a) Marginalia: ah, eh, oh, uh, ajá, mm,mh,um,ts.

(b) Citas y estilo directo: “ ” (sin espacio).

*Otras cuestiones*

(a) Los nombres de persona se reemplazan por una letra mayúscula, incluidos los participantes en turnos.

(b) El anonimato debe preservarse, así que se emplearán iniciales u otros subterfugios para lugares o cualquier información que permitiera la identificación del informante.

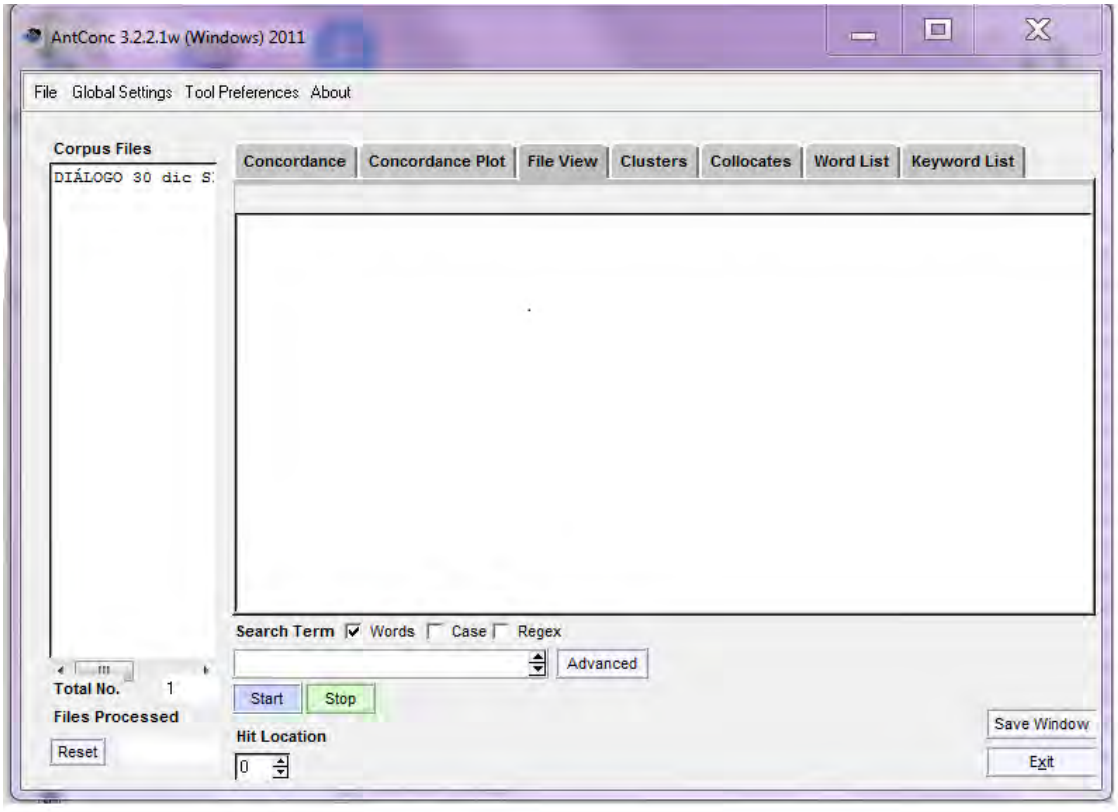
(c) Los turnos de habla se numerarán, pero hasta que la trascripción esté completamente revisada.

(d) Para traslapes por ser lo más habitual; ( ) para observaciones diversas –así se escriben las acotaciones; < > para texto repuesto, es normal en edición de textos.

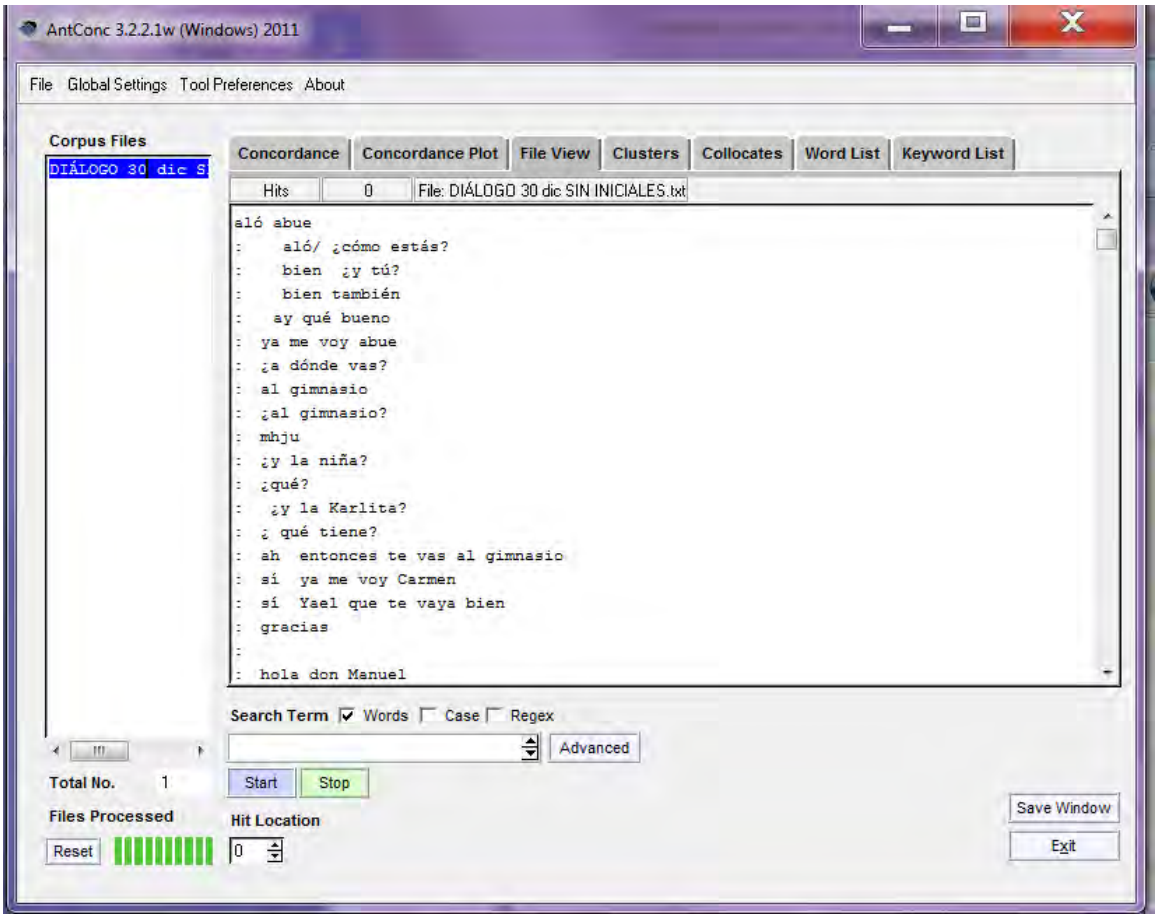
# APÉNDICE II

## EJEMPLO DE USO DEL PROGRAMA *ANTCONC*

El primer paso para obtener los resultados de *Ant Conc* 3.2.2.1w es guardar el texto que se desea como "texto sin formato", posteriormente se abre el programa *Ant Conc* 3.2.2.1w y se busca el documento en *File* / *Open Files* y se selecciona el texto deseado previamente guardado sin formato.

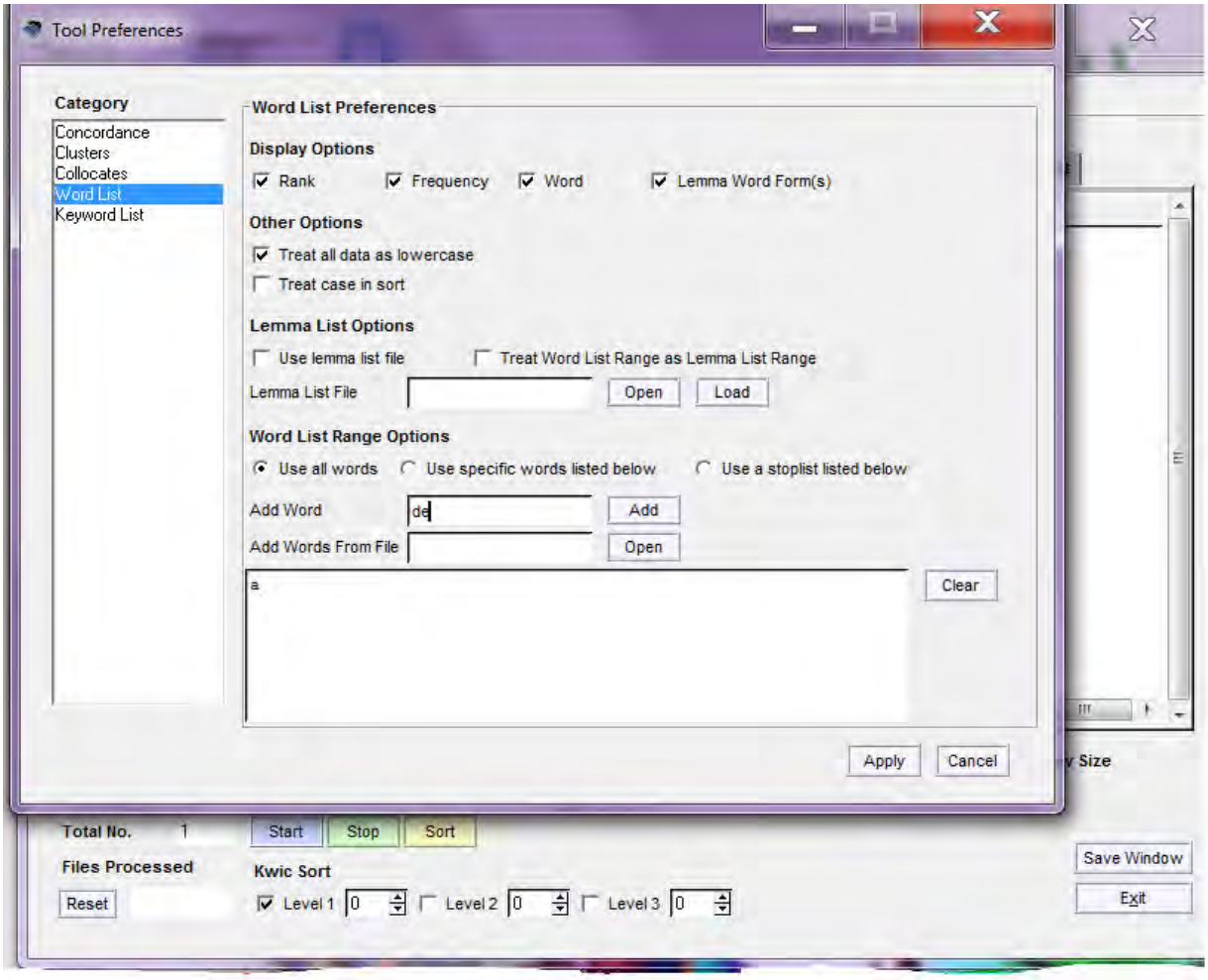


El texto podrá verse si se selecciona  en las pestañas *File View* y se le da *clic* al nombre del archivo que se encuentra debajo de *Corpus Files*.



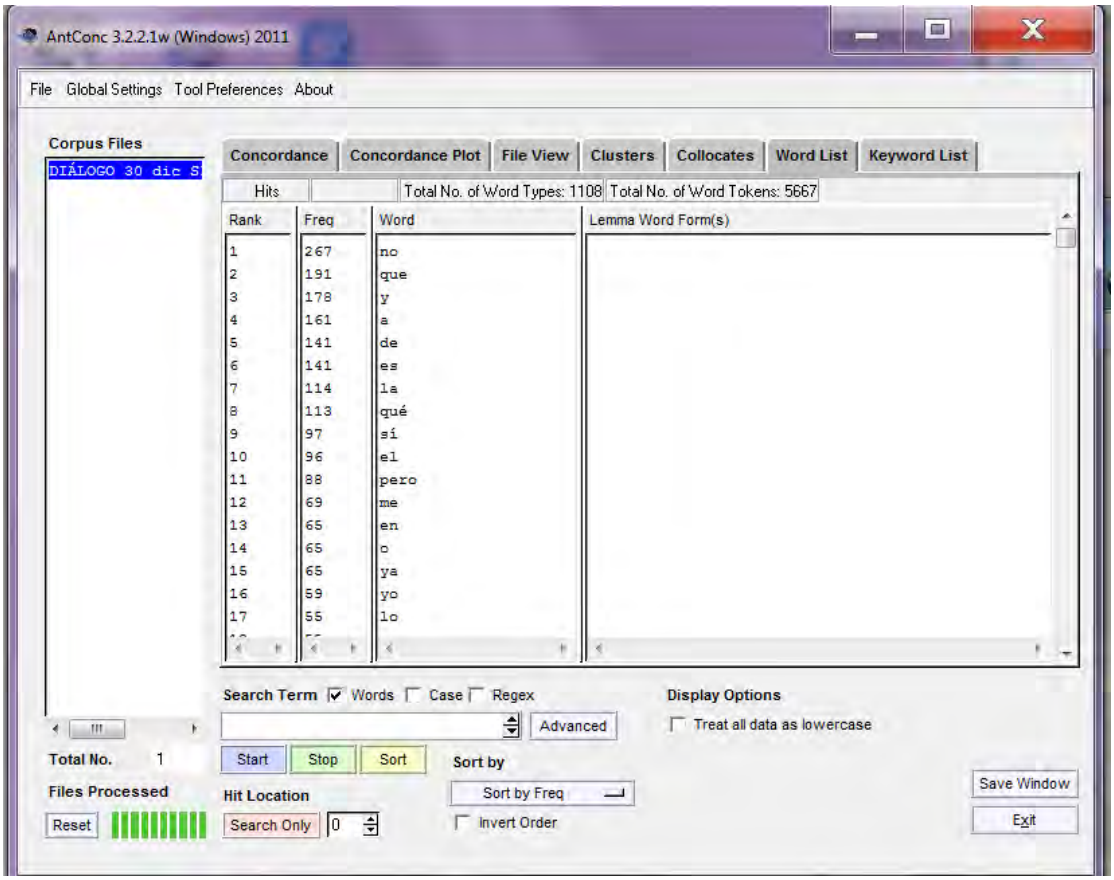

Después, para indicar las órdenes a seguir por el programa, se selecciona *Tool Preferences,* posteriormente en *Category*  se selecciona *Word List,* en *Other Options* se selecciona *Trate all data as lowercase* para que no haga distinción entre mayúsculas y minúsculas, si se necesita ignorar alguna palabra en el conteo se selecciona *Use a stoplist listed bellow* que se encuentra en la sección de *Word List*

*Rage*, esto se ubica dentro de la misma ventana y en *Add Word* se introducen una por una las palabras que no se desean oprimiendo *Add* para que el programa las registre, una vez registradas todas las palabras se oprime *Apply*




Después se selecciona la pestaña *Word List* y se oprime el botón *Start*, a continuación se verá el listado de las palabras del texto con su frecuencia y en la

sección de arriba el total de tipos de palabras o *Types* y el total de palabras o *Tokens*.[2]




Este programa junto con *Wave Pad Sound Editor* facilitó la obtención de datos para este proyecto. Ambos programas fueron herramientas que contribuyeron a procesar los datos para su posterior análisis.

[2] Para una explicación con más detalle ingresar a la página: http://www.slideshare.net/laurarp/gua-antconc-321-presentation.