UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
LICENCIATURA EN LENGUA Y LITERATURAS HISPÁNICAS

EL ESTUDIO DE LOS DIPTONGOS DEL ESPAÑOL DE MÉXICO PARA SU APLICACIÓN EN UN RECONOCEDOR DE HABLA

TESIS QUE, PARA OBTENER EL TÍTULO DE
LICENCIADA EN LENGUA Y LITERATURAS HISPÁNICAS,
PRESENTA

FERNANDA LÓPEZ ESCOBEDO

Asesores: Lic. Javier Octavio Cuétara Priede
Dr. Luis Alberto Pineda Cortés

Ciudad de México, 2004

ÍNDICE

ÍNDICE DE CUADROS

ÍNDICE DE IMÁGENES

1. INTRODUCCIÓN

En los últimos años, ha habido un gran interés por desarrollar programas informáticos que permitan a la computadora procesar el lenguaje natural. De esta gran área se desprenden distintos campos de trabajo, como por ejemplo, el procesamiento de textos y las tecnologías del habla. Ésta última se divide en síntesis de habla y reconocimiento de habla; la primera tiene el propósito de transformar automáticamente un texto escrito en habla, mientras que la segunda tiene como finalidad la tarea inversa.

Este trabajo tiene la intención de proponer una nueva transcripción fonética computacional de los diptongos para mejorar la calidad de un programa de reconocimiento de habla. El proceso que se lleva a cabo parte de un corpus oral que se transcribe tanto fonética como fonológicamente. Posteriormente, esto sirve para entrenar a la computadora por medio de modelos acústicos que permiten el reconocimiento de una señal sonora.

Durante la transcripción fonética del corpus oral se delimita, con ayuda de imágenes espectográficas, el inicio y el final de cada uno de los alófonos que lo componen. Así, a cada alófono le corresponde una etiqueta diferente. Sin embargo, el caso de los diptongos y de las sinalefas que forman diptongo genera ciertos problemas ya que es muy difícil delimitar el inicio y el final de cada una de las vocales que lo conforman. Hasta ahora, la solución más fácil y rápida ha sido cortar por la mitad el segmento que corresponde al diptongo, y transcribir así una y otra vocal. No obstante, en transcripciones de este tipo el reconocedor podría toparse con ciertas dificultades si las etiquetas fonéticas

computacionales de un diptongo no corresponden exactamente con el limite entre cada una de sus vocales.

A partir de este problema se pensó en una posible solución que, además de contribuir a una mejor calidad en el reconocimiento, fuera fácil y rápida de transcribir. Por lo tanto, en este trabajo se propone la transcripción fonética de los diptongos de un corpus oral que tendrá la finalidad de crear un sistema de reconocimiento de habla como una unidad; es decir, asignar una sola etiqueta fonética computacional al segmento que forma el diptongo.

De esta manera, en el primer capítulo del presente trabajo se consideró necesario comenzar con una introducción al estudio del procesamiento de lenguaje natural. Posteriormente, se explican con más detalle las tecnologías del habla y se presentará un esquema que permite comprender el proceso que se lleva a cabo en el reconocimiento de habla.

El siguiente capítulo mostrará el proceso que se lleva a cabo en la transcripción fonética de un corpus oral, el Corpus DIME. Para esto, será necesario hablar de las herramientas principales que intervienen en este proceso, como lo son las imágenes espectrográficas, los alfabetos computacionales y las técnicas para crear un corpus oral.

Con el propósito de tener una base teórica lingüística, en el capítulo cuatro se dará la definición de cada uno de los fenómenos fonéticos que presentan dos vocales contiguas: diptongos, hiatos, sinalefas y diptongación de hiatos. Además, como introducción a estos fenómenos se proporcionará una breve clasificación de las vocales. Para concluir este capítulo se contrastarán las ideas de dos autores, Trubetzkoy y Alarcos, en cuanto a que el primero considera los diptongos como monofonemáticos, mientras que el segundo se inclina par otorgarles un carácter bifonemático.

Finalmente, en el último capítulo, se analizará la aparición de estos cuatro fenómenos que presentan dos vocales contiguas en un corpus oral del español de México. Comenzará por explicar el Proyecto DIME, en el que este trabajo se inscribe, así como el Corpus DIME, que fue el objeto de estudio de esta tesis para el análisis de los datos. Posteriormente, se mostrarán los datos que se obtuvieron y los resultados finales que de ellos se desprenden.

De esta manera,

1. **queda establecida una propuesta para transcribir computacionalmente a los diptongos como monofonemáticos con miras a las tecnologías del habla,**
2. **y se abre, dentro del propio Proyecto DIME, la posibilidad de un trabajo futuro que se encargue de emplear esta información para el trabajo práctico.**

Parece acertado recordar que, "el tiempo para el desarrollo de tecnología del lenguaje humano es particularmente oportuno, ya que el mundo se moviliza en el desarrollo de la supercarretera de la información, que será el soporte del futuro crecimiento económico. Las tecnologías del lenguaje humano desempeñarán un rol central en proveer una interfaz que cambie drásticamente el paradigma de comunicación hombre-máquina, de *programación* a *conversación.* Esto permitirá a los usuarios accesar eficientemente, procesar, manipular, y absorber una gran cantidad de información" (Olivier, 1999:18).

2. PROCESAMIENTO DE HABLA, TEXTO Y LENGUAJE NATURAL

Debido a que los resultados de este trabajo impactarán en la arquitectura de un reconocedor de habla para el español de México, es necesario tener presentes las primeras propuestas que se hicieron al respecto. Para ello, se mencionarán algunos de los autores que fueron fundamentales en el desarrollo de la moderna ciencia computacional. Se señalará la diferencia entre los tres campos de trabajo en los que se puede dividir el estudio del lenguaje natural, los cuales son el procesamiento de lenguaje natural, el procesamiento de textos y las tecnologías del habla.

Posteriormente, se profundizará en ésta última, la cual incluye tanto reconocimiento como síntesis de habla. En este apartado se hablará de la diferencia que existe entre una y otra y de la importancia de ambas para las nuevas tecnologías del habla. Finalmente, se presentará un breve esquema de la arquitectura de un reconocedor de habla, con la finalidad de mostrar dónde impacta la propuesta de este trabajo.

2.1. Antecedentes

La lingüística computacional es una interdisciplina donde conviven computólogos, ingenieros y lingüistas. Por ende, las investigaciones relacionadas con esta fecunda área de investigación comprenden conocimientos computacionales y lingüísticos.

En el aspecto computacional, esta disciplina es el resultado de décadas de investigación. Desde los años cincuenta cuando Turing (*apud.* Jurafsky *et al.* 2000:10) se

planteó la posibilidad de modelar el pensamiento por medio de las computadoras, y esbozó el programa de investigación en inteligencia artificial, se iniciaron las distintas líneas de estudio que en la actualidad han llevado a la aparición de áreas como la lingüística computacional y el procesamiento del lenguaje natural.

Dentro de estos años de trabajo, se pueden mencionar distintos autores que fueron claves para el desarrollo de la moderna ciencia computacional. Entre ellos se encuentra Chomsky (1956, *apud.* Jurafsky *et al.* 2000:11) que, además de su trabajo en la gramática generativa, es uno de los principales autores de la teoría del lenguaje formal y de los autómatas, fundamentos básicos de la ciencia computacional. Así como también Elwood Shannon quien, entre otras cosas, desarrolló la teoría matemática de la información.

Con estas teorías como antecedente, la lingüística computacional surgió como un área interesada en desarrollar programas informáticos capaces de generar lenguaje natural. Al respecto Jurafsky *et al.* (2000) han considerado tres campos que se desprenden del estudio del procesamiento del habla y del lenguaje natural: el procesamiento de lenguaje natural, la lingüística computacional y la síntesis y el reconocimiento de habla.

No obstante, desde una perspectiva más amplia, a partir de esta opinión y de las necesidades que han surgido en los últimos años -como es la aparición de la red electrónica mundial y la necesidad de extraer únicamente los datos que interesan en un momento dado y de desechar los que son muy poco relevantes- es posible considerar tres áreas principales en este campo de investigación: el procesamiento del lenguaje natural, el procesamiento de texto y las tecnologías del habla. El primero tiene por objetivo entender el significado y la intención de un texto, el segundo trabaja con el texto para facilitar su manejo sin importar el significado del mismo, y el tercero tiene como objetivo el tratamiento del habla para producir sistemas de síntesis y reconocimiento de habla.

Por un lado, el procesamiento de texto tiene distintas aplicaciones, como son la recuperación de información, la traducción automática, los resúmenes automáticos y la generación automática de textos, entre otros. Por ejemplo, uno de los problemas a los que se enfrenta la recuperación de información es, ante una consulta en la red, acceder únicamente a la información que el usuario desea obtener, desechando toda información que sea irrelevante a la búsqueda inicial. Como respuesta a este problema, se han creado analizadores morfológicos, léxicos computacionales y léxicos multilingües, entre otros, que se utilizan a manera de recursos para crear técnicas en la recuperación de información; ésta se centra en dos aspectos fundamentales: el filtrado de los datos y el multilingüismo. "En el primer caso se trata de diseñar sistemas de recuperación de información que recuperen sólo los documentos que interesan al usuario, que eviten el exceso de «ruido» -es decir, la recuperación de textos irrelevantes- y que no ignoren ningún texto que pueda ser de interés. En lo que se refiere al segundo aspecto, se hace necesario contar con recursos que permitan el acceso multilingüe a los datos, los cuales, a su vez, son también multilingües" (Llisterri *et al.*, 2002:19).

Por otro lado, el área de procesamiento del lenguaje natural tiene como objetivo fundamental obtener y procesar una representación del significado del texto. Para ello, se utilizan desde recursos morfológicos y semánticos hasta sintácticos y pragmáticos, con el fin de generar un dialogo que permita que el sistema responda al usuario de la forma más correcta.

Finalmente, el campo de las tecnologías del habla, donde se centra la presente investigación. Las tecnologías del habla tienen como objetivo principal "facilitar la interacción oral con los sistemas informáticos, complementando o substituyendo los métodos tradicionales como el teclado o la pantalla" (Llisterri *et al.*, 2002:19). Se pueden

distinguir dos tipos principales de tecnologías del habla: la síntesis, que posibilita una salida de habla, y el reconocimiento, que posibilitan una entrada de información oral.

A continuación se profundiza en las tecnologías del habla, por ser el área donde se desarrolla este trabajo.

2.2. Tecnologías del habla

Las primeras propuestas que se hacen en el campo de las tecnologías del habla se remontan a por lo menos tres décadas. Llisterri (2003:256) menciona que es a principios de los años setenta cuando el control de la producción del habla por parte de una computadora empieza a ser posible y cuando surgen técnicas digitales que permiten manipular las señales acústicas del habla. En el transcurso de esta década, se perfeccionaron los sintetizadores de habla -o sistemas de conversión de texto en habla- y los sistemas que reconocían palabras aisladas. Posteriormente, en la década de los años ochenta, los mayores adelantos -más allá del simple reconocimiento de palabras aisladas- se dieron dentro del reconocimiento de habla continua, al eliminarse la necesidad de introducir pausas entre cada palabra para que la computadora reconociera un enunciado. En los años noventa se disponía ya de productos comerciales para la síntesis de habla más naturales que los de épocas anteriores. En el reconocimiento de habla, el mayor avance se situó en el reconocimiento de grandes vocabularios sin limitaciones.

Actualmente, las tecnologías del habla han representado un gran interés para algunas compañías y centros de investigación, pues permiten la interacción humano-máquina en tiempo real. Los servicios de telefonía son algunas de ellas, ya que mediante una conversación con un sistema informático acceden a una serie de funciones, como

reservación de boletos y obtención de información sobre horarios de transportes, entre otras; por otra parte, estos sistemas también pueden ser aplicados en apoyo a los discapacitados que no pueden usar un teclado para operar una computadora, e incluso para crear ambientes de alta tecnología, como son los edificios y los automóviles inteligentes.

No obstante, de acuerdo con Olivier (1999), a pesar de las ventajas que ofrecen las tecnologías del habla -como las de permitir al humano tener las manos y ojos libres, y además la posibilidad de localizarse en cualquier sitio- aún existen limitaciones, como lograr respuestas en tiempo real, resolver las variaciones de ruido y logras sistemas de reconocimiento independientes del locutor.

Como se mencionó en el primer apartado (§2.1), el trabajo que se ha realizado a lo largo de todas estas décadas tiene la característica de ser una interdisciplina. Crear un sistema de habla no es una tarea que se pueda hacer individualmente; es necesaria la intervención de distintas áreas, donde cada una de ellas tiene una labor específica para la creación del sistema. A este respecto, Acero (1995) opina que en este campo interdisciplinario hay cupo para ingenieros, computólogos y lingüistas: “Unlike the speech sciences, whose main goal is to gain a better understanding of the speech production and generation process, speech technology’s main goal is to build systems. Therefore, a linguist willing to pursue a career path in speech technology has to be a practical person and a team player” (Acero, 1995).

Con el fin de crear un modo eficaz en el surgimiento de las tecnologías del habla, Llisterri *et al*. (1998) mencionan que existen dos módulos:

1. **La síntesis de habla: que es la generación automática del habla a partir de una representación simbólica.**
2. **El reconocimiento de habla: que es la conversión del habla en una representación simbólica.**

Ambas tecnologías se integran en sistemas que permiten crear un diálogo entre la persona y la computadora.

La síntesis de habla es la encargada de transformar automáticamente un texto escrito en una realización sonora. Para que esto suceda, el texto debe estar en formato electrónico. Uno de los problemas que se presentan es el de los signos que no corresponden a una palabra ortográfica pronunciable, como es el caso de abreviaturas, siglas, números, etc. Por ejemplo, siglas como "UNAM", que se refieren a *Universidad Nacional Autónoma de Mëxico* o como "S. XII" que se refiere a *siglo doce*. Es necesario interpretarlos y obtener un texto legible para la computadora. "Es decir, para que un texto pueda ser oralizado apropiadamente, debe presentarse como una unidad lingüísticamente coherente y cohesionada, sin ambigüedades involuntarias. La calidad de un conversor mejora si se incorporan las técnicas y herramientas desarrolladas dentro del ámbito del procesamiento del lenguaje natural, así como sus resultados, en el modulo de análisis lingüístico" (Llisterri *et al*, 2003).

Existen distintos procedimientos para conseguir que un sistema computacional sea capaz de generar un mensaje oral. El más sencillo es el sistema que reproduce un mensaje previamente grabado; el número de oraciones capaz de producir depende de las introducidas anteriormente por medio de la grabación.

Un sistema más complejo es el que genera un número ilimitado de oraciones. Para lograrlo, se utilizan las principales características del lenguaje humano con el fin de crear un número infinito de enunciados a partir de un número finito de unidades, como pueden ser alófonos, fonemas, sílabas, palabras y frases. "Cuanto menor sea la unidad, más reducido será el inventario que necesitaremos pero será más fácil alcanzar una buena

concatenación o unión entre las unidades que permita un habla sintetizada con la mayor naturalidad posible, sin saltos ni interrupciones" (Llisterri, 2003:260).

Lo más común para los sistemas de síntesis que se basan en la concatenación de unidades son las semisílabas y los difonemas. Lo que se hace en este tipo de sintetizadores es almacenar las propiedades acústicas de las unidades primeramente grabadas por un hablante y reconstruirlas cuando sea necesaria la generación de un mensaje. Para ello, existen diversos procedimientos como: la síntesis por LPC (*Linear Predictive Coding*), la síntesis de los formantes, la síntesis por reglas y la síntesis articulatoria.

La síntesis por reglas, por ejemplo, se utiliza cuando se tratan unidades básicas como los fonemas y alófonos: "Se establece un conjunto de reglas que determinan las características acústicas de cada unidad y la forma de concatenarlas" (Llisterri, 2003:264). Ésta es una técnica que se basa en los conocimientos fonéticos obtenidos a partir del análisis del habla; por el contrario, la síntesis articulatoria, en lugar de partir de la información acústica, controla al sintetizador por medio de la posición de los articuladores.

Un sistema de reconocimiento de habla realiza la tarea inversa, pues convierte la señal sonora del habla en una representación simbólica. Este es el sistema que se quiere crear en el proyecto del que forma parte este trabajo, Proyecto DIME, el cual será abordado en su oportunidad (§5.1).

Existen distintos tipos de reconocedores, dependiendo de los criterios a los que se atiende. Puede haber reconocedores según el número de locutores que reconocen o según el tamaño del vocabulario que reconocen, entre otros.

Al describir la arquitectura de un reconocedor de habla, Llisterri y sus colaboradores (2003) mencionan que estos sistemas aprenden de un extenso corpus de habla que, al enfrentarse con un nuevo enunciado, comparen los datos con los previamente obtenidos en

la etapa de aprendizaje; así, serán capaces de llevar a cabo el reconocimiento. Por ello, "el corpus de entrenamiento debe contener la mayor variedad posible de realizaciones para que puedan crearse los modelos de cada una de las unidades, reflejando, entre otros elementos, la variación individual de las voces, los distintos acentos debidos a factores geográficos o sociolingüísticos y las diferencias en la velocidad de elocución que puedan darse entre hablantes"(Llisterri *et al.*, 2003).

Para llevar a cabo la etapa de aprendizaje de un reconocedor de habla, es necesario crear la transcripción, tanto fonética como ortográfica del texto, alineándola temporalmente con la señal sonora. Posteriormente, se incorpora un diccionario que contiene las palabras que serán utilizadas por el sistema, transcritas fonéticamente, incluyendo sus posibles variantes de pronunciación. Así, a partir de una señal sonora, el programa lleva a cabo el reconocimiento de patrones que le permite seleccionar el modelo que más se asemeja a la entrada. Esto se puede realizar a través de diferentes modelos de ingeniería, como las redes neuronales y los modelos ocultos de Markov, o un modelo híbrido de ambos.

Se obtiene así que, la diferencia principal entre un reconocedor de habla y un sintetizador es la siguiente: "In an ASR[1] system the main objective is to improve the recognition accuracy and usability. In a TTS[2] system, the major objective is to improve speech quality. Any work that is not headed along those lines will not benefit system's performance" (Acero, 1995).

En el siguiente apartado se profundizará en la arquitectura de un reconocedor de habla, detallando el proceso que se lleva a cabo y las etapas que intervienen en él. Esto, con

[1] ASR: *automatic speech recognition*: reconocimiento automático del habla.
[2] TTS: *text to speech*: síntesis de habla.

el fin de mostrar dónde impacta la propuesta que presenta esta tesis en cuanto al etiquetado fonético computacional de los diptongos del español, como una unidad.

2.3. Arquitectura de un reconocedor de habla

Para el proceso de reconocimiento de habla Tapias (2002:191) menciona la necesidad de cuatro tipos de información:

1. **los modelos acústicos que permiten identificar los sonidos, pues proporcionan la información sobre las propiedades y características de los mismos.**
2. **el diccionario donde están el conjunto de sonidos que forma cada palabra del vocabulario.**
3. **los modelos del lenguaje que contienen la información de cómo se deben combinar las palabras para formar frases.**
4. **los modelos predictivos que, en el caso de los sistemas de diálogo, el reconocedor dispone de predicciones sobre el contenido de la siguiente frase que pronunciará el locutor.**

Se tiene así que la arquitectura de un reconocedor de habla se puede ilustrar como se presenta en la imagen 1.

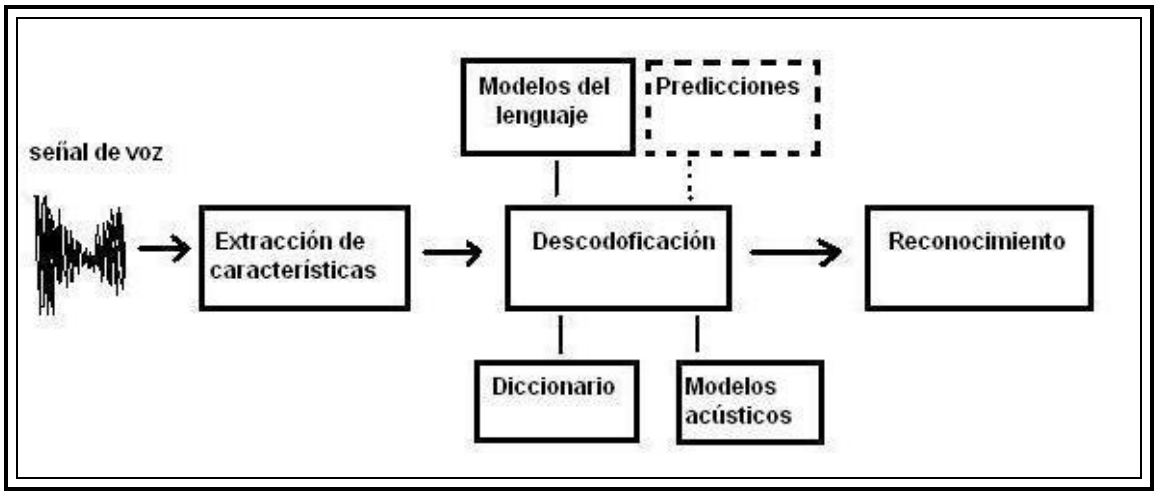

Imagen 1. Arquitectura de un reconocedor de habla

Al inicio del proceso, se genera una señal de voz la cual, en la extracción de características, se convierte a un formato que pueda ser correlacionado con información lingüística. Esta información se pasa al descodificador, el cual se apoya en los modelos del lenguaje, en el diccionario, en los modelos acústicos y en la predicción en el caso de un sistema de diálogo.

Los modelos acústicos son los que se obtienen en el proceso de entrenamiento a partir de un conjunto finito de muchas muestras de habla. Estos modelos pueden ser a nivel de alófono, de fonema, de sílaba o de palabra. El diccionario contiene las palabras que constituyen el corpus utilizado para crear los modelos acústicos. En éste está presente el conjunto de sonidos que forma cada una de las palabras del corpus. Los modelos del lenguaje son las palabras que pueden seguir a cada una de las palabras utilizadas en el corpus. Finalmente, en un reconocedor de diálogo es necesaria la presencia de predicciones que "en función del estado de la conversación, toman la decisión más adecuada y hacen una predicción sobre la siguiente interacción con el usuario" (Tapias 2002:192). Para ello, es necesaria la utilización de un analizador semántico que extraiga el significado de la frase.

A partir de la información que se tiene con los modelos acústicos, el diccionario y los modelos del lenguaje, el programa lleva a cabo el reconocimiento de la señal sonora.

En el caso específico del Proyecto DIME, se utilizó un corpus en el dominio de cocinas (Villaseñor *et al.*, 2001). Los modelos acústicos se complementarán a partir de la transcripción fonética de los alófonos del español de México, pronunciadas por 15 sujetos diferentes. El entrenamiento se está haciendo actualmente con un repertorio finito de estos alófonos. Si la propuesta de etiquetar los diptongos como una sola unidad fuera válida, éstos tendrían que incluirse dentro de este repertorio y, por lo tanto, el número de modelos acústicos aumentaría conforme al número de diptongos que hay en la lengua.

Consecuentemente, en el diccionario del reconocedor las palabras que estuvieran constituidas por un diptongo deberían contener esta información.

Lo anterior para el caso de los diptongos; sin embargo, para las sinalefas, como veremos, se presenta otro tipo de problema. Al tener una palabra que termine con una vocal seguida de otra que empiece con una vocal, se puede formar un diptongo. Consiguientemente, en el diccionario de pronunciación del Corpus DIME, además de incluir las palabras con diptongo, se podría, en otro trabajo futuro, considerar la posibilidad de incluir las palabras que pronunciadas en habla continua se fusionan para formar un diptongo entre ellas y constituir una sola, sobre todo en el caso de ciertas combinaciones de palabras que son frecuentes en el corpus. Por ejemplo en la oración *necesito un fregadero*, la sinalefa, que en habla continua se formaría en la vocal final de *necesito*, y la primera vocal de *un*: [... ne se sí tou9n ...][3] se representaría como un diptongo y, por lo tanto, quedaría como una sola palabra: *necesitoun*. Este trabajo futuro debería considerar al léxico restringido tomando en cuenta el número de palabras por las que esta formado el corpus. Además, esto repercutiría en los modelos del lenguaje, ya que las dos palabras en sinalefa formando diptongo se tomarían como una unidad y se tendrían que agregar las combinaciones de palabras que podrían preceder a ésta.

El gran problema que presenta la etiquetación fonética computacional de diptongos y las palabras que en sinalefa forman un diptongo, podría verse solucionado con la propuesta de la presente investigación. El reconocedor no se presentaría ante el problema de dos palabras que al entrar en contacto pierden su autonomía, se convierten en una sola y, por ende, el reconocimiento del inicio del segundo elemento resulta una tarea muy difícil para

[3] Cabe mencionar que las transcripciones fonéticas que aparecen en este trabajo utilizan el alfabeto de la *Revista de Filología Española* (1915).

el programa. Además, el trabajo de transcripción fonética se vería beneficiado en cuanto a la optimización del factor tiempo para etiquetar los diptongos.

En el capítulo siguiente se hablará de las herramientas (incluyendo alfabetos) que se utilizan para el proceso de transcripción fonética, así como la descripción del proceso en sí.

3. FONÉTICA Y FONOLOGÍA COMPUTACIONAL

En la construcción de un reconocedor de habla, el lingüista juega un papel muy importante, por lo que es necesario hablar de los aspectos donde la intervención lingüística es fundamental. Para esto, se hará una introducción donde se abordará el término fonética instrumental, así como también se mostrará la herramienta que es de gran utilidad para el análisis acústico de los sonidos, el espectrograma. Posteriormente, se mencionarán algunos de los alfabetos fonéticos mas utilizados hoy en día; del mismo modo, se hará alusión a ciertos alfabetos computacionales que se derivan de éstos.

A continuación se hablará de las distintas técnicas que existen para crear un corpus oral, profundizando en los experimentos "Mago de Oz". Finalmente, se retomarán los aspectos mencionados, para detallar el proceso que se lleva a cabo en una transcripción o etiquetación fonética.

3.1. Fonética instrumental

La fonética constituye hoy en día un fértil campo interdisciplinario, pues se relaciona con otras disciplinas que requieren y/o hacen uso del análisis de los sonidos del habla; algunas de estas áreas son la informática, la ingeniería, la computación y la neurología, entre otras.

Existen, además, distintas áreas de estudio dentro de la fonética. Entre ellas, se encuentra la fonética instrumental o fonética experimental, que se ocupa de métodos de registro y medición instrumentales. Una definición la da Solé (1985), quien dice que "la

fonética instrumental tiene algunos puntos en común con la fonética y la fonología experimental. En primer lugar, es producto de la influencia de las ciencias naturales sobre el desarrollo de la lingüística [...] En segundo lugar, la fonética y la fonología experimental también se sirven de instrumentos de registro y medición para la observación objetiva de los fenómenos lingüísticos [...] En tercer lugar, los estudios fisiológicos y acústicos de los fonetistas del siglo XIX y su desarrollo de métodos y técnicas instrumentales de análisis han sido una base para la metodología experimental" (Solé, 1985).

La fonética instrumental ha tenido un gran desarrollo en los últimos años gracias al reciente progreso tecnológico, el cual ha proporcionado el equipo necesario para medir los fenómenos acústicos. De esto se han favorecido disciplinas como las tecnologías del habla, donde el reconocimiento y la síntesis del habla han sido su interés principal. Los estudios que se realizan del español en el campo de la dialectología también se han beneficiado del uso de este tipo de herramientas que permiten proporcionar datos empíricos.

A continuación se hará un repaso de los primeros instrumentos que se han utilizado para medir la onda sonora del habla humana y los adelantos que ha habido hasta nuestros días.

3.1.1. Herramientas (espectrógrafo)

Desde el siglo XVIII, se llevaron a cabo intentos por construir un aparato que pudiera medir la onda sonora. Por un lado, Carl Ludwing inventó el quimógrafo, mientras que Hermann Helmholtz trató de reproducir los sonidos vocálicos mediante un aparato, y formuló la teoría de la resonancia. Posteriormente, Abbé Rousselot fue quien "definitivamente llevase hasta donde alcanzaba la técnica del momento los conocimientos de la fonética acústica,

basándose en instrumentos, aún muy toscos, que permitían aproximarse a la verdad física del sonido emitido y aprovechado en la comunicación humana" (Martínez Celdrán, 1984:91). Finalmente, en el siglo XIX surgió el espectrógrafo o sonógrafo que, descrito por Quilis (1981:84) tiene la función de descomponer automáticamente la onda sonora compleja en cada uno de sus componentes integrantes y, de esta manera, proporciona todos los datos de los sonidos del habla que nos interesa conocer, como la frecuencia formántica de cada fonema y su duración, entre muchas otras funciones.

Actualmente, es muy fácil contar con un espectrógrafo en una computadora personal, como es el caso de la herramienta *SpeechView*, del *Center for Spoken Language Undestanding* del *Oregon Graduate Institute* (CSLU/OGI). Ésta puede ser instalada de manera gratuita en cualquier equipo y nos proporciona el análisis espectrográfico de cualquier onda sonora que se quiera analizar. Todo ello es el resultado del avance tecnológico que ha habido en los últimos años, pues en el siglo pasado el espectrógrafo era una aparato de grandes dimensiones y difícil de manejar.

Para entender mejor la representación espectrográfica de un sonido, en la imagen 2 se presenta el espectrograma de la vocal *a*, producida por un hablante femenino, con el programa *SpeechView*:

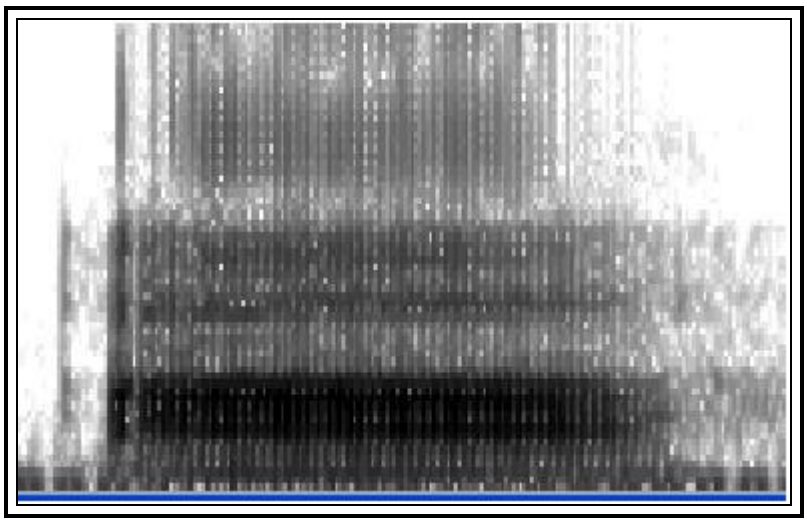

Imagen 2. Espectrograma de la vocal *a*

En esta imagen, se pueden observar dos líneas bien definidas denominadas formantes. "La excitación producida por la glotis se extiende a las cavidades supraglóticas, que, al actuar como filtros, estructuran la señal acústica. Desde la glotis hasta los labios se pueden considerar una serie de cavidades resonantes que, a través de la función de transferencia o, lo que es lo mismo, de la función del filtrado del conducto vocal, originan los llamados *formantes*. Un formante de la onda acústica del lenguaje es, por tanto, un máximo de la función de transferencia del conducto vocal" (Quilis, 1981:138).

Tanto Alarcos (1950) como Quilis (1985) coinciden en la idea de que:

1. **El formante 1 de una vocal indica el grado de abertura del conducto vocal: cuanto más alta es la frecuencia del F1, más grande es la abertura total de la cavidad, y a la inversa.**
2. **El formante 2 señala la longitud de la cavidad bucal: cuanto más alta es la frecuencia del F2, menor es la longitud de la cavidad bucal, y a la inversa.**

Para entender mejor esta diferencia, en la imagen 3 se muestran los formantes 1 y 2, de cada una de las vocales del español.

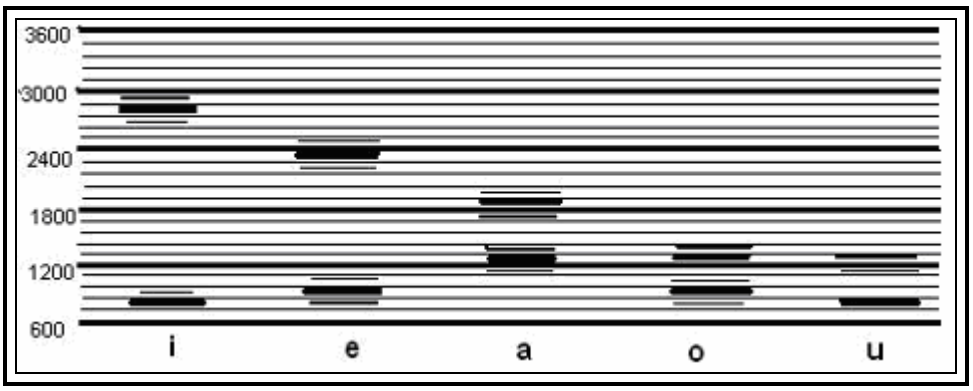

Imagen 3. Formantes de las cinco vocales del español[4]

Identificar los formantes de cada vocal nos ayuda a realizar una mejor lectura del espectrograma; sin embargo, también es importante tomar en cuenta la duración: "al introducir la variable tiempo en la representación de la onda sonora, el espectrograma permite identificar la trayectoria de los formantes y por tanto obtener información sobre las transiciones de un sonido a otro. No obstante, determinar el punto en que se inicia el cambio de frecuencia y el punto en el que se estabiliza de nuevo la trayectoria del formante, es decir, determinar los límites de la transición, es especialmente difícil cuando las muestras de habla proceden de una situación comunicativa informal, caracterizada por una relajación en la pronunciación" (Llisterri *et al.*, 1999). Este problema se ve reflejado en el momento del etiquetado fonético,[5] pues a pesar de que los formantes de las vocales y algunas consonantes como las líquidas y las nasales se observan claramente, en el caso del resto de las consonantes lo que ofrece un espectrógrafo es la modificación que éstas ejercen

[4] En este trabajo únicamente se muestran los formantes para el caso de las 5 vocales del español; sin embargo, en la *Fonética acústica de la lengua española*, de Quilis (1981), puede observarse la tabla de formantes para los alófonos de las vocales hecha por Delattre (*apud*. Quilis, 1981).

[5] En la introducción ya se ha mencionado brevemente lo que es el etiquetado fonético, empero en el último apartado de este capítulo se explicara con mayor detalle el proceso que se lleva a cabo en el etiquetado fonético de un corpus oral.

sobre los formantes de las vocales vecinas. En estos casos, la lectura del espectrograma resulta un tanto confusa, y es difícil encontrar la frontera exacta entre los sonidos.

Fuera de estas dificultades, el uso de esta herramienta ha beneficiado en gran medida el desarrollo de las tecnologías del habla. Un reconocedor de habla es necesario que sea entrenado con modelos del lenguaje, y estos modelos se crean a partir de un corpus oral que se etiqueta fonéticamente. Gracias al espectrógrafo, el fonetista es capaz de delimitar los fonemas que conforman la onda sonora, y así crear una serie de transcripciones o etiquetas que serán los modelos con los cuales el reconocedor será entrenado.

El espectrógrafo ha sido una herramienta primordial dentro de la fonética y la fonología computacional, así como también la creación de alfabetos que permitan la transcripción de los sonidos que se observan en un espectrograma.

3.2. Alfabetos fonéticos

Los alfabetos fonéticos que más se utilizan en la actualidad en nuestro país, son el Alfabeto Fonético Internacional (AFI) o *International Phonetic Alphabet* (IPA), y el alfabeto de la *Revista de Filología Española* (*RFE*). Este último con mucho mayor aceptación por los filólogos hispanistas.

La Asociación Fonética Internacional fue fundada en 1886 por un grupo de maestros de lingüística. La idea de establecer un alfabeto fonético que pudiera ser aplicado a todas las lenguas fue de Otto Jespersen: “The alphabet of the *Association Phonétique Internationale* is an alphabet on romanic basis designed primarily to meet practical linguistic needs, such as putting on record the phonetic or phonemic structure of languages,

furnishing learners of foreign languages with phonetic transcriptions to assist them in acquiring the pronunciation, and working out romanic orthographies for languages written in other systems or for languages hitherto unwritten." (IPA, 1949:1). Sin embargo, la misma asociación opina que este alfabeto no es perfecto, pues es necesario incorporar nuevos signos para sonidos que se vayan descubriendo.

En 1999 la Asociación crea el *Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet,* el cual es un nuevo trabajo que retoma y revisa lo hecho en 1949.

La importancia de este nuevo libro es que toma en cuenta el avance tecnológico que se ha desarrollado en los análisis acústicos y propone un nuevo enfoque del alfabeto. "The new *Handbook* is intented to be a reference work not only for language teachers and phoneticians interested in the sounds of different languages, but also for speech technologists, speech pathologists, theorical phonologists, and others" (IPA, 1999:1).

Por otro lado, la *Revista de Filología Española* (*RFE*) fue fundada en 1914, y en seguida se planteó la necesidad de tener un alfabeto propio que sirviese para todo tipo de investigación fonética en el habla hispánica. Finalmente, en 1915 lo publicó. A partir de esa fecha, la investigación filológica y lingüística en España y en América Latina ha utilizado este alfabeto, hoy en día, principalmente en nuestro continente.

Martínez Celdrán (1984:147) opina que las diferencias que existen entre uno y otro son que el de la *RFE* hace un mayor uso de los signos diacríticos, mientras que el de la AFI prefiere letras nuevas e introduce gran cantidad de signos griegos. Estas diferencias se explican porque los objetivos del alfabeto de la AFI consistieron en sus inicios en el uso exclusivo de éste para la enseñanza de lenguas extranjeras, mientras que el de *RFE*, para la

investigación filológica y lingüística hispánica. Sin embargo, actualmente es mucho mayor el uso de AFI para la investigación lingüística en general.

Al respecto, De la Mota *et al*. (1995) opinan lo siguiente: "como es sabido, la *Revista de Filología Española* propone en 1915 un alfabeto fonético que se difunde sobre todo dentro de la tradición románica hispánica, pero que en la actualidad es ya poco utilizado. Por otra parte el Alfabeto Fonético Internacional aparece en 1889, y sus objetivos se publican en 1949 en los *Principles of the International Phonetic Association.* Este alfabeto, revisado por última vez en 1999, posee, en cambio, ámbito internacional" (De la Mota, 1995).

Actualmente, en México el alfabeto que se enseña más comúnmente es el de la *RFE*; sin embargo, debería también enseñarse el de la IPA pues prácticamente todos los estudios internacionales de fonética y fonología utilizan este último.

3.2.1. Alfabetos computacionales

Existe otro tipo de alfabetos fonéticos que se utilizan actualmente en investigación lingüística: los alfabetos fonéticos computacionales; sin embargo, presentan varios problemas, como lo refiere Llisterri: "las convenciones permiten la posibilidad de una doble representación de los sonidos y los diacríticos adoptados no contemplan todas las caracterizaciones fonéticas. A estos problemas se suman las limitaciones del medio informático, allí donde la técnica condiciona los medios de representación lingüística; la proliferación de diacríticos complica la codificación informática del alfabeto" (Llisterri *et al*., 1999).

Como respuesta a estos problemas, surge un alfabeto fonético electrónico SAMPA (*Speech Assessment Methods Phonetic Alphabet*), que es considerado como una versión informática de AFI. Este alfabeto fue desarrollado bajo el proyecto ESPRIT, compuesto por un grupo de fonetistas internacionales, y fue usado, en primera instancia, en lenguas de la comunidad europea: "SAMPA basically consists of a mapping of symbols of the International Phonetic Alphabet onto ASCII[6] codes" (Blaheta, Ms.). Además, este alfabeto fue realizado con la colaboración de investigadores de distintos países, y los símbolos fueron desarrollados consultando hablantes nativos de las distintas lenguas a las que fue aplicado.

El problema de este alfabeto computacional (SAMPA) fue que no establecía una codificación para todos los símbolos del AFI, por lo que posteriormente se propuso un nuevo alfabeto denominado X-SAMPA. Llisterri (1997) dice que en esta ampliación de SAMPA se prevén equivalencias en códigos ASCII para la totalidad de símbolos del AFI, incluyendo diacríticos y marcas tonales.

Sin embargo, tanto SAMPA como X-SAMPA son alfabetos que sólo se aplicaron a las lenguas europeas. Ante esta dificultad, Hieronymus (1994 y 1997) propuso un nuevo alfabeto: Worldbet, en el cual se representan una gran variedad de alófonos para todas las lenguas del mundo, desde las europeas, hasta las asiáticas y africanas. Para ello, incluyó símbolos en códigos ASCII que visualmente tienen una semejanza con los del AFI. Un ejemplo es el caso de la consonante alveolar fricativa dentalizada, representada en AFI como [s5]. Lo que hace Worldbet es buscar una equivalencia con el símbolo de dentalización utilizado en AFI y lo resuelve de la siguiente manera s[. "Worldbet is an

[6] Las letras ASCII representan las siglas en inglés de la frase *american standard code for information interchange*; es decir, códigos que presenten una uniformidad en el intercambio de información electrónica.

attempt to have a phonetic alphabet wich covers all of the world's languages in a systematic fashion. It is an ASCII version of the IPA plus a number of symbols wich were found useful in database labeling, which are not currently on the official IPA set. This list of extra symbols may grow with time until all of the important phenomena have a coherent symbol representation" (Hieronymus, 1994).

Es interesante observar que Worldbet contiene una lista de los diptongos que aparecen en lenguas europeas; sin embargo, Hieronymus (1994) opina que no es el total de ellos, y dice que "new diphthongs can be constructed by concatinating the two vowel symbols corresponding to the beginning and ending vowels of the diphthong".

Posteriormente, en un nuevo artículo en 1997, Hieronymus retoma el problema antes mencionado y nos dice que "the present set in this document is the subset of Worldbet which is useful for phonemic labeling of languages, and thus most suitable for speech recognition and synthesis. With the exception of diphthongs, this set should be capable of covering the sounds of all of the languages in the World" (Hieronymus, 1997). Por lo tanto, Hieronymus en 1997 ya plantea la posibilidad de crear sonidos nuevos que representen un diptongo.

Finalmente, un nuevo alfabeto fonético computacional, llamado Mexbet, se ha creado para los fonemas y alófonos del español de México. Está basado en Worldbet e incluye los símbolos para el español de México. Además, si se quisiera, este alfabeto podría ser utilizado para representar las variaciones observadas en algunos dialectos, así como para la representación del español Sudamericano.

Los autores originales de este alfabeto son Uraga y Pineda (2000) quienes al respecto dicen, "we propose a phonetic alphabet based on Worldbet in which symbols for all Mexican and South American Spanish phonemes are included; this alphabet is also

complemented with symbols for the phonemes of the Castillian Spanish. Symbols for the allophonic variations observed in such dialects are also provided for. We refer to this phonetic alphabet as Mexbet".

En el Proyecto DIME se utiliza este alfabeto, en una nueva y más especializada versión (Cuétara, 2004), para etiquetar fonéticamente el corpus, ya que es una representación del español de México y por lo tanto es de gran utilidad recurrir a los símbolos que propone Mexbet. En el cuadro 1 se muestra la última versión del alfabeto Mexbet.

Consonantes	Labiales	Labiodental	Dentales	Alveolares	Palatales	Velares
Oclusivo sordo	p		t		k_j	k
Oclusivo sonoro	b		d			g
Africado sordo					tS	
Africado sonoro					dZ	
Fricativo sordo		f	s_[	s		x
Fricativo sonoro	V		D	z	Z	G
Nasales	m			n	n~	N
Vibrantes			n_[	r(/ r		
Laterales				l		
Vocales				Anteriores	Media	Posteriores
Paravocales				j		w
Cerradas				i		U
Medias				e		o
Medias abiertas				E		O
Abierta				a_j	a	a_2

Cuadro 1. Tabla del alfabeto fonético Mexbet

Tenemos así que, a los ojos de este trabajo, el alfabeto en código ASCII más completo y actual es Worldbet. No obstante, si el estudio que se lleva a cabo -como es el caso de este trabajo- se refiere específicamente al español de México, es definitivamente imprescindible utilizar Mexbet.

Hasta el momento se han hablado de los alfabetos con los que se puede realizar una trascripción fonética y de la herramienta principal que se utiliza para delimitar los sonidos que conforman un dialogo. No obstante, estas dos herramientas son empleadas a partir de que se tiene una señal sonora, la cual puede ser un corpus oral que tenga una finalidad específica.

3.3. Corpus orales

Un corpus oral se puede crear de distintas formas, algunas de ellas serían presentar al hablante una lista de palabras u oraciones que debe leer, un cuestionario que obligue al hablante a utilizar ciertas palabras para el interés de la investigación, o simplemente un dialogo libre entre dos hablantes.

En el caso del Corpus DIME, se realizó de la última manera, donde uno de los hablantes simulaba ser una computadora y el otro fungía en el papel de usuario. Esto se hizo con el interés de observar la interacción que podría existir entre un sistema y un usuario; además, se determinó un dominio específico, que en este caso fue el diseño de una cocina: “In an experimental situation, the message can be completely pre-planned or scripted, partially planned, or completely unplanned or unscripted. However, it is worth mentioning that this feature may change if the speaker or the experimenter is considered. In semi-directed interviews or in specific tasks the researcher has planned the contents of the corpus he wants to obtain, although the speaker might not be aware of it” (Aguilar *et al*., 1994).

Además, una de las principales finalidades de crear un reconocedor de habla, o cualquier programa que implique la interacción de humano-máquina, es la de entablar una conversación en lenguaje natural entre una persona y un sistema computacional lo mas semejante a la que se establecería entre dos humanos.

Con el objetivo de lograr este tipo de diálogo, Dahlbäck *et al*. (1993) utilizan el método que se conoce como "Wizard of Oz", o los experimentos "Mago de Oz".

3.3.1. Los experimentos "Mago de Oz"

Los experimentos "Mago de Oz" tienen como objetivo simular un diálogo en lenguaje natural que surja de la interacción entre un sistema y un hablante. Para ello, es necesario que el usuario crea que está interactuando con una computadora cuando en realidad está hablando con otra persona que funge el papel de sistema. "Studies where subjects are told that they are interacting with a computer system through a natural-language interface, though in fact they are not. Instead the interaction is mediated by a human operator, the wizard, with the consequence that the subject can be given more freedom of expression, or be constrained in more systematic ways" (Dahlbäck, 1993). Este fue el método que se llevó a cabo para crear el Corpus DIME, con la diferencia de que, en este caso, el usuario sabía que con el que estaba interactuando también era una persona.

La primera diferencia que encontramos entre una computadora y un humano es el lenguaje y la manera de expresarse de cada uno. Mientras que el primero es rígido, rápido y sin errores, el segundo es flexible, lento y comete errores, es por ello que, para que el usuario se imagine que en realidad sí esta hablando con una computadora, es necesario cuidar todos estos detalles.

Dahlbäck *et al*. (1993) proponen tres fenómenos importantes que se deben considerar: "the background system, the task given to subjects, and the wizard's guidelines and tools.

1. **The background system should be simulated or fully implemented. A shaky prototype will only reveal that system's limitations and will not provide useful data. Furthermore, the system should allow for a minimum of mixed-initiative dialogue. A system-directed background system will give a dialogue which is not varied enough. However, if the purpose is to collect data from the use of a particular application, or for the development of an interface for a particular system, then that application will determine the interaction.**
2. **The task given to subjects must be reasonably open, i.e. have the form of a scenario. Retrieving information from a database system and putting it together for a specific purpose can be fruitful [...]**
3. **Finally, we have the simulation environment and guidelines for the wizard. The simulation experiment must be studied in detail, from pilot experiments, before the real simulations are carried out. This information is used to provide knowledge to the wizard on how to act in various situations that may be encountered" (Dahlbäck *et al*, 1993).**

Además, los experimentos "Mago de Oz" se deben llevar a cabo en un laboratorio que tenga dos cuartos para que en uno se sitúe al Mago y en otro al usuario. Así fue como se realizó en el Proyecto DIME, en el cual se logró el objetivo que era el de observar un diálogo en lenguaje natural entre una supuesta computadora y un humano.[7]

Al tener un corpus oral enfocado en el interés de la investigación, un alfabeto computacional y un espectrograma que permite delimitar los sonidos, se puede llevar a cabo la transcripción o etiquetado fonético.

3.4. Transcripción de corpus orales

Actualmente, la transcripción fonética electrónica ha tenido sus aplicaciones más interesantes en el desarrollo de sistemas en tecnologías del habla, como en el caso de la

[7] En el capítulo 5 (§5.2) se explicará con mayor detalle el proceso que se llevó a cabo en la creación del Corpus DIME.

síntesis y reconocimiento de habla. En una transcripción fonética se "determinan los puntos de inicio y fin de cada segmento individual, lo que se conoce como etiquetado de una señal, y se requiere para poder crear un sistema de reconocimiento de voz" (Olivier, 1999:6).

El propósito del etiquetado fonético computacional es el de precisar la información lingüística que contiene una onda sonora. Así, se alinean los límites ortográficos y fonéticos manualmente, utilizando una herramienta que permita observar la onda sonora y colocar etiquetas. Llisterri (1997) opina que "para cada nivel de representación suele establecerse un conjunto de «etiquetas» que se asocian a un determinado fragmento del corpus -un segmento sonoro, una unidad prosódica, una palabra, etc.- y definen sus propiedades. Las etiquetas propias de un nivel fonético de representación corresponden a las características articulatorias o acústicas de los sonidos del habla, mientras que, por ejemplo, las etiquetas de un nivel de representación morfosintáctico describen propiedades morfológicas y léxicas de las palabras. El etiquetado constituye, por tanto, un enriquecimiento del corpus mediante información adicional introducida por el investigador en función de sus objetivos y, lo que es más importante, de su interpretación lingüística de los materiales recogidos" (Llisterri, 1997).

Como ya se mencionó en el primer apartado de este capítulo (§3.1.1), una de las herramientas que nos permite observar la onda sonora y colocar etiquetas es con la que se trabaja en el Corpus DIME, el *SpeechViewer* del CSLU/OGI.

Lander en su "*The CSLU labeling guide*", señaló que "speech data al CSLU are transcribed at two levels: orthographic and broad phonetic. We produce non-time aligned orthographic to provide quick access to the content of an utterance[8]. Some orthographic

[8] *Utterance*, se refiere a las unidades mínimas en que se puede segmentar un diálogo. El término utilizado en español es el de elocuciones.

transcriptions contain markers for word boundaries, to support access and retrieval at the lexical level. Time aligned phonetic transcriptions give a more detailed representation of the utterance to enable phonetic and phonemic analysis" (Lander, 1997:1).

La barra de herramientas del *SpeechView* se presenta en la imagen 4:

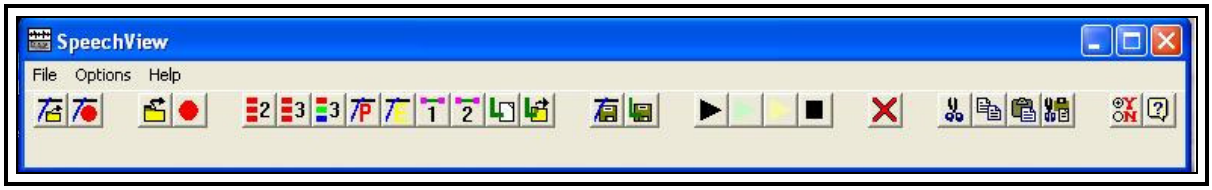

Imagen 4. Barra de herramientas del programa *SpeechView*, del CSLU/OGI

Esta herramienta permite observar una onda sonora guardada anteriormente en un archivo, grabar a un hablante, y guardar el archivo de habla; observar los espectrogramas en blanco y negro en segunda y tercera dimensión, y a color en tercera dimensión; observar la curva entonativa; crear etiquetas y guardarlas; reproducir toda la onda sonora o por segmentos, y editar la onda sonora, entre otras funciones.

Para llevar a cabo la transcripción fonética computacional, es necesario tener un conocimiento sobre lectura de espectrogramas, pues de esta manera es como se localizan los bordes entre fonemas y se asignan las etiquetas para cada uno de ellos.

La primera transcripción suele ser la de nivel ortográfico (Llisterri, 1997), que en determinado tipo de corpus se acompaña de una transcripción fonética o fonológica. Esto es porque existen corpus orales donde no es factible realizar una transcripción fonética completa, pues existe un número elevado de horas de grabación.

En la imagen 5 se muestra un ejemplo de una transcripción ortográfica.

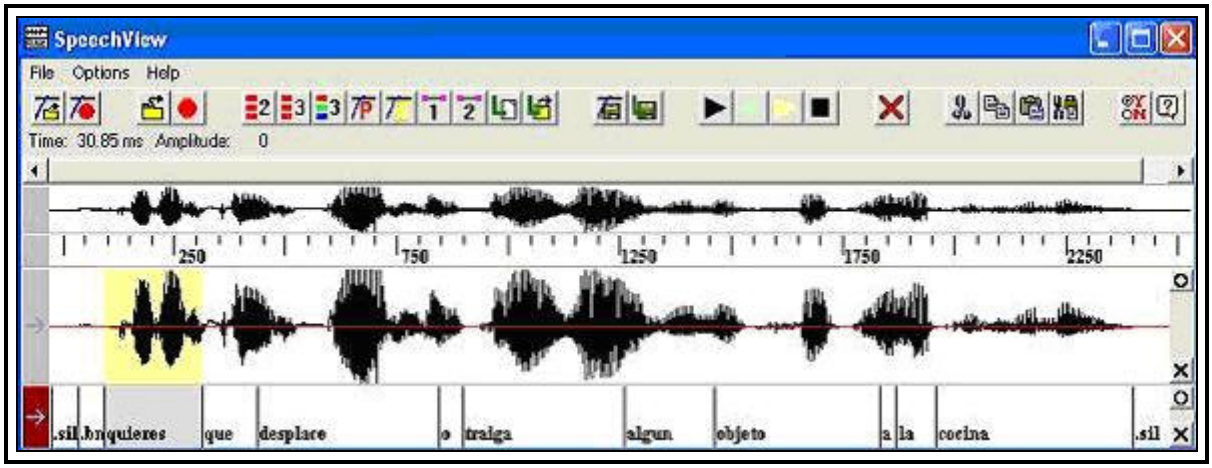

Imagen 5. Transcripción ortográfica de un fragmento del Corpus DIME utilizando el programa *SpeechView*

Posteriormente, se pasa a la transcripción fonética del corpus, la cual debe estar alineada temporalmente con la señal sonora y con la transcripción ortográfica. Además de las etiquetas para los fonemas, existen unas más que describe los ruidos ajenos al habla que se está estudiando. Lander señala "keep in mind that these labels are used in two types of transcriptions, non-time aligned and time aligned. At the none-time aligned level the labels appear in pointy brackets (e.g. <bn>); at the time aligned level they appear with a preceding period (e.g. .bn)" (Lander, 1997:19) y proporciona una lista de sonidos, de los cuales, en el caso del Corpus DIME, únicamente se utilizan los sonidos .sil (*silence*-silencio), y .bn (*back noise*-ruido de fondo) (Cuétara, 2004).

Un ejemplo de la transcripción fonética del Corpus DIME se muestra en la imagen 6.

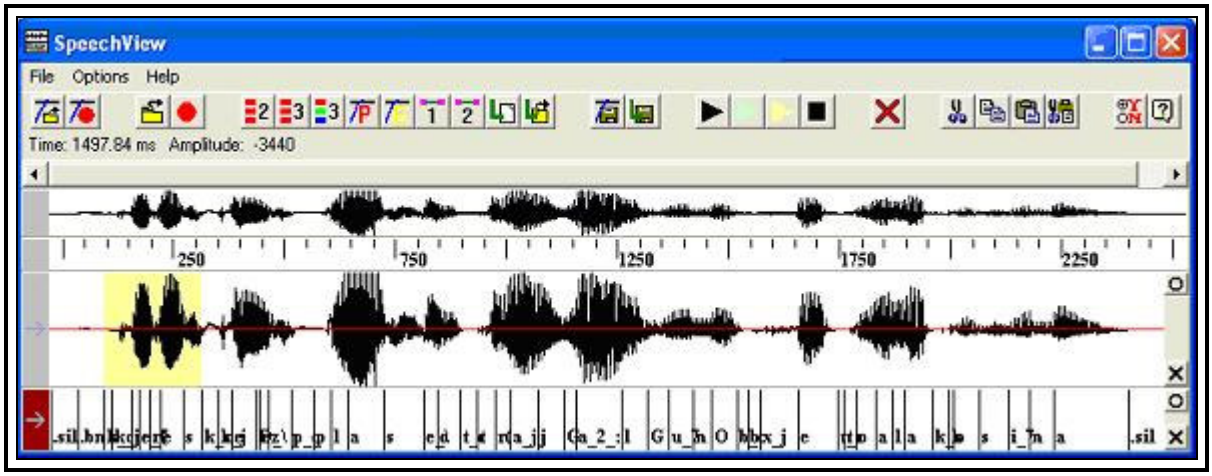

Imagen 6. Transcripción fonética de un fragmento del Corpus DIME utilizando el programa *SpeechView*

El grupo Tlatoa (2000), el equipo de investigación sobre tecnologías del habla de la Universidad de las Américas, propone los pasos que se deben llevar a cabo para etiquetar un corpus:

1. **"Create .files for each speaker you wish to label […] The .files file should contain the name of the .wrd file as the last filename on each line, although this is not standard practice in the CSLU Toolkit code [...]**
2. **Generate non-time-aligned text transcriptions. These transcriptions are contained in text files with the extension .txt. They consist of the words spoken [...]**
3. **These transcriptions are contained in .wrd files.**
   They consist of the words spoken and the beginning and end times of each word [...]
4. **Adjust by hand the word-level labels [...]**
5. **Trim long silences from the wave files [...]**
6. **Generate time-aligned phonetic labels automatically from the word-level labels. The goal here is to do the same for phonemes as we did for words [...]**
7. **Adjust by hand the phonetic-level labels".**

El proceso que se llevó a cabo en el Corpus DIME fue muy similar; lo único que difiere es la parte de etiquetado automático, o *time alignment*, que corresponde a los pasos 6 y 7, pues la transcripción sólo se realizó de manera manual, por lo que los resultados obtenidos son mucho más precisos. No obstante, la desventaja que existe en un etiquetado manual es que se consume mucho más tiempo y se requiere de habilidad y conocimientos específicos

para identificar la porción de señal que corresponde al fonema, razones por las cuales surge la necesidad de crear un sistema capaz de realizar el proceso de manera automática.[9]

El tiempo que se requiere para transcribir fonéticamente una oración de 15 a 20 palabras es de aproximadamente 20 minutos. Si se considera que un corpus puede estar conformado por 5000 oraciones, el tiempo que se lleva a cabo para realizar el etiquetado fonético es muchísimo: de tal suerte, contar con una herramienta que proporcione una transcripción fonética automática resulta de gran utilidad.

Dentro del Proyecto DIME, se realizó un transcriptor automático para el español de México: *TranscríbEMex* (Cuétara, 2004) (Pineda *et al*., 2004) el cual, al escribir en una ventana una palabra, frase u oración, proporciona la transcripción fonética, fonológica y la división silábica de la misma. Este programa ofrece rapidez para etiquetar grandes corpus orales, asimismo, favorece la uniformidad en las transcripciones fonéticas y elimina ambigüedades. A pesar de que se debe ajustar manualmente la alineación fonética del corpus, ha sido una herramienta muy útil durante la etapa de transcripción fonética computacional del Corpus DIME.

Para concluir este capítulo, es necesario hablar de lo que sucede en la transcripción fonética de los diptongos. En un corpus, donde el habla es continua, es más difícil encontrar las fronteras que existen entre los fonemas, ya que éstos son influenciados por el fonema anterior y el que le sigue. Lander opina así, y dice, "phones are not always signaled by discrete, non-overlapping regions in the waveform. In continuous speech, coarticulation, deletion, and elision cause phone boundaries to overlap. Therefore, because true boundaries

[9] Alejandra Olivier en 1999 hizo una tesis en torno a este problema que lleva por nombre *Evaluación de métodos de determinación automática de una transcripción fonética* y propone un posible transcriptor automático.

do not actually exist in many cases, care must be taken to follow convention in order to ensure consistency" (Lander, 1997:43).

Si esto sucede con la mayoría de los fonemas presentes en habla continua, es todavía más grave en el caso de los diptongos, donde aún en habla pausada es difícil o casi imposible identificar los limites entre cada uno de los segmentos que los constituyen.

Como solución a este problema, Lander propone lo siguiente: "after a semivowel or vowel, it can be practically impossible to determine the exact onset of a vowel. To be consistent, we have chosen to place the boundary in the middle of the transition period" (Lander, 1997:50). Esta podría ser una solución; sin embargo, al dividir un diptongo por la mitad se corre el riesgo de que no siempre sea el límite exacto entre uno y otro fonema y, por lo tanto, al momento de entrenar al sistema éste se toparía con graves problemas.

Por todo ello, esta tesis propone una mejor solución: etiquetar los diptongos como un solo segmento. Esto evitaría, por un lado, la dificultosa tarea de encontrar los límites de las vocales existentes en un diptongo y, por otro, solucionar las limitaciones que podría encontrar un reconocedor de habla a la hora de toparse con este tipo de sonidos agrupados.

En la imagen 7 se contrasta la transcripción fonética computacional de los diptongos que hasta ahora se ha utilizado, con la que se propone en este trabajo.

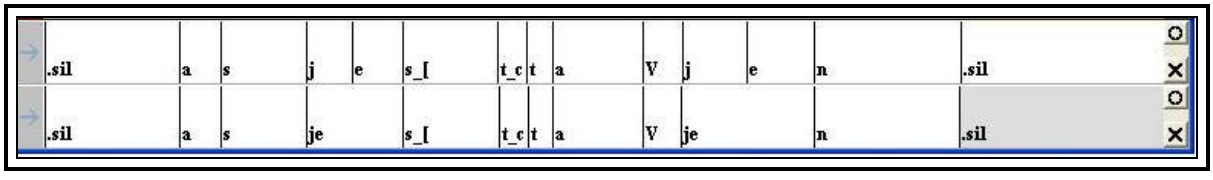

Imagen 7. Transcripción fonética de los diptongos y nueva propuesta de transcripción fonética para los diptongos

En este capítulo se describió el proceso que se lleva a cabo en la transcripción fonética de un corpus oral, para lo que fue necesario hablar de los aspectos que intervienen y que son necesarios para que esto se pueda realizar eficientemente. Así, se habló de la fonética instrumental, del espectrograma, de los alfabetos fonéticos y computacionales, y de la creación de corpus orales. Principalmente, nos detuvimos en los aspectos relacionados con el etiquetado fonético computacional, y la dificultad que representa el análisis acústico de los diptongos. En el siguiente capítulo se dará la definición de cada uno de los fenómenos fonéticos que presentan dos vocales contiguas: diptongos, hiatos, sinalefas y diptongación de hiatos.

4. DIPTONGOS E HIATOS DEL ESPAÑOL DE MÉXICO

Si el propósito principal de este trabajo es proponer una nueva transcripción fonética computacional para el caso de los diptongos del español de México, es necesario estudiar lo que las teorías lingüísticas dicen acerca de ellos, así como también lo que sucede en los casos en los que se encuentran dos vocales contiguas y no forman diptongo, como el de los hiatos.

Este capítulo comenzará por introducir las características principales de las vocales del español, para posteriormente profundizar en los distintos fenómenos que de ellas se desprenden. El primero será el caso de los diptongos, incluyendo el de las sinalefas que forman diptongo, seguido del de los hiatos, para concluir con el de la diptongación de hiatos. Se dará una definición de cada uno de estos fenómenos, así como una muestra de su estudio a partir de la fonética instrumental con el apoyo de imágenes espectrográficas para su mejor explicación. Finalmente, se expondrán las opiniones de dos importantes lingüistas, Trubetzkoy y Alarcos, sobre el carácter monofonemático o bifonemático, de los diptongos.

4.1. Clasificación de las vocales

Del repertorio de los 22 fonemas que existen en el español de México, cinco son vocálicos. Como Vaquero menciona (1996:11), el español mantiene en Hispanoamérica, al igual que

en España, un sistema fonológico de cinco unidades; la clasificación de éstas se basa en dos factores:

1. **El grado de abertura, "que se identifica con el modo de articulación" (vocal abierta: /a/, medias: /e, o/ y cerradas: /i, u/).**
2. **La posición de la lengua, "que determina el lugar de articulación" (central: /a/, anteriores o palatales: /i, e/ y posteriores o velares: /o, u/).]**

Por lo tanto, las vocales, según su modo y punto de articulación, pueden representarse en un triángulo ideado por el alemán Hellwag en 1781. "En dicho triángulo, dispuesto de manera invertida, los vértices superiores van ocupados por la i (vértice palatal) y por la u (vértice velar), correspondiendo el vértice inferior a la vocal a. Entre la a y la i se colocan la e y las demás vocales palatales intermedias, y entre la a y la u, las velares" (Navarro Tomás, 1918:37). Esto se aprecia en la imagen 8.

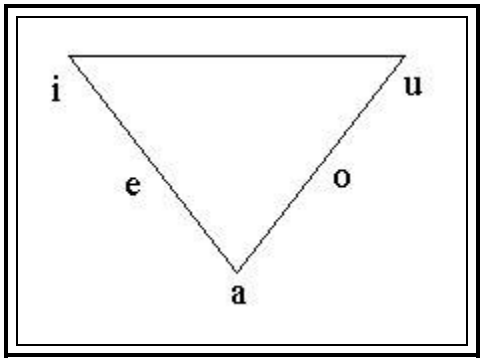

Imagen 8. Triángulo vocálico para el español

Al dividir el triangulo de manera vertical, se tiene que, según su punto de articulación, las primeras vocales /e/, /i/, forman la serie de vocales palatales; /a/ es una vocal central, y /o/ y /u/, forman la serie de vocales velares. Dividiendo el triángulo de manera horizontal, se tiene que, según su modo de articulación, /i/ y /u/ corresponden a

vocales cerradas o de abertura mínima, /e/ y /o/ son vocales de abertura media y, finalmente, /a/ es la vocal abierta o de máxima abertura.

Además de los rasgos distintivos de los fonemas vocálicos (modo y lugar de articulación), es importante tomar en cuenta las distintas realizaciones que existen de cada uno de ellos. "Un alófono se define por medio de todos sus rasgos, tanto de aquellos que son distintivos cuando funciona como unidad fonológica, como los que no los son" (Quilis, 1985:67). Por ende, encontramos que cada fonema vocálico puede llegar a tener un número infinito de alófonos; sin embargo, en la práctica son sólo unos cuantos los que se repiten frecuentemente. Los alófonos de cada uno de los fonemas vocálicos que considera el *Esbozo* (1973) se aprecian en el cuadro 1.

Fonema		**Alófonos**
/i/	**Alto anterior**	**[i, i7, j, i9]**
/e/	**Medio anterior**	**[e, e7]**
/a/	**Central bajo**	**[a, a*f*]**
/o/	**Medio posterior redondeado**	**[o, o7]**
/u/	**Alto posterior redondeado**	**[u, u7, w, u9]**

Cuadro 2. Alófonos de las vocales según el *Esbozo de una nueva gramática de la lengua española* (1973)

Para el caso de la vocal anterior alta, además de su forma prototípica, como en [si], existen 3 alófonos posibles dentro del español: uno abierto, como en [r#í7sa], la semiconsonante anterior [kjéres], y la semivocal anterior [péi9ne]. De la vocal media anterior encontramos la realización abierta [temé7r]. De la *a* se tiene el alófono cerrado en [a*f*ĉa]. De la vocal alta posterior encontramos -al igual que para *i*- tres alófonos: el abierto, en palabras como [r#ú7ta], la semiconsonante en [kwéte] y la semivocal en [eu9rópa].

Finalmente, para la vocal media posterior encontramos su realización abierta en palabras como [ko7r#o7].

Estos fenómenos alofónicos, que se combinan con el rasgo vocálico -y que no necesariamente son exclusivos de las vocales- pueden deberse a distintas causas, algunas de ellas pueden resultar de la posición de la lengua y los labios a la hora de emitir el fonema, por el grado de abertura, o por la posición que éste ocupa en la sílaba, entre otros.

Además de los rasgos distintivos que presentan cada una de las vocales, desde el punto de vista de la fonética acústica "no se ha llegado aún a establecer con precisión si las vocales, aparte del tono diferencial correspondiente a la abertura y calidad de cada una de ellas, muestran también alguna diferencia de altura, dentro de la línea melódica de la palabra, como efecto de la distinta articulación y timbre de dichos sonidos" (Navarro Tomás,1944:21).

Después de conocer las características de las vocales del español, es necesario mencionar la importancia de éstas en la lengua. Entre muchos otros factores, su importancia radica en que pueden constituir por sí solas una sílaba, por ejemplo la preposición /a/ o la conjunción /y/, a diferencia de los fonemas consonánticos que carecen de esta propiedad en el español. Incluso, como menciona Alarcos (1950:145), las vocales del español pueden formar palabras entre sí -solas, aisladamente o combinadas-, y da como ejemplo la preposición *a*, la forma verbal *he*, las conjunciones *y*, *o* y *u*, e, incluso, las formas *ahí*, *oí* y *huía*.

Ya que en una sílaba del español debe existir la presencia de una vocal, *El esbozo de una nueva gramática de la lengua española* (RAE, 1973) define al elemento vocálico de la sílaba como cima (Ci). La cima puede ser simple o compuesta. Es simple la que contiene

una sola vocal, compuesta la que está formada por un grupo de dos o tres vocales. Por otra parte, cada una de las palabras con significado de nuestra lengua estará conformada con una vocal como mínimo. Consecuentemente, el porcentaje de aparición de las cinco vocales debe ser alto.

Para constatar si esta afirmación es correcta, es necesario comparar los datos que se han obtenido en estudios sobre frecuencia de aparición de fonemas en la lengua española, y lo que encontramos es que, en prácticamente todos los autores que veremos, existe una distribución más o menos equivalente entre las 17 consonantes y las cinco vocales. Éstas últimas representan casi el 50 % del total de fonemas.

Emilio Pérez Hernán en su artículo *Frecuencias de fonemas* (2003), dice que las vocales representan un 46.23% del total de los fonemas y las consonantes el 53.77% restante. Sin embargo, no es el único que ha hecho estudios al respecto, Alarcos (1965), Navarro Tomás (1946), Guirao y García Jurado (1993), Llisterri y Mariño (1993), Quilis (1981), Cuétara (2004) entre otros, también han estudiado la frecuencia de fonemas del español y coinciden, no exactamente pero con muy poca discrepancia, con los datos obtenido por Emilio Pérez Hernán.

Se concluye así que, a pesar de que las vocales representan menos de una cuarta parte del total de fonemas, pues únicamente representan 5 de los 22 fonemas del español, tienen un porcentaje de aparición casi equivalente al 50 %.

4.1.1. Imagenes espectrográficas de las vocales

Como se mencionó en el capítulo anterior, la fonética instrumental utiliza el espectrógrafo para descomponer la onda sonora en cada uno de sus componentes integrantes. Utilizando

este aparato para percibir los componentes de las vocales, se observan dos formantes bien definidos; por un lado el primer formante (F1), que retomando lo que ya se ha dicho muestra el grado de abertura, y por otro lado el segundo formante (F2), que muestra la longitud de la cavidad oral.

A propósito de las vocales del español de México, Madrid y Marín (2001) resaltan el hecho de que la frecuencia de los formantes es siempre más alta en las vocales producidas por mujeres que en las de los hombres, lo cual se debe a la diferencia en la longitud del tracto vocálico, que, en promedio, es de 15 a 20% más corto en las mujeres que en los hombres. Esto hace que las frecuencias de una misma vocal varíen dependiendo del sexo del hablante.

En la imagen 9 se muestra tanto el oscilograma como el espectrograma de las cinco vocales del español. Se puede observar la altura de los formantes de cada una de ellas.

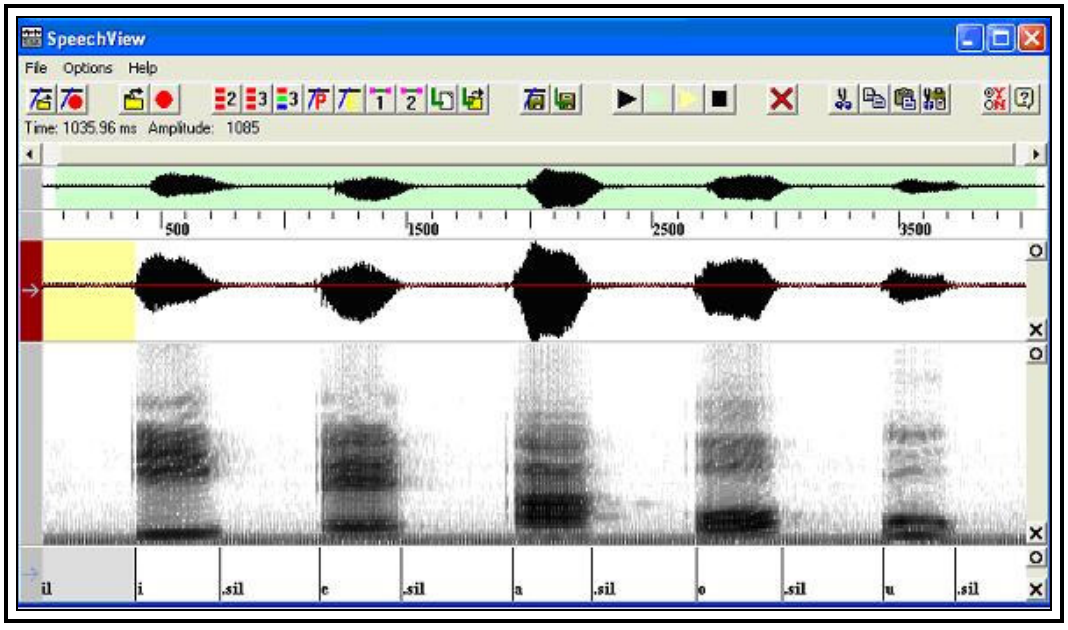

Imagen 9. Las cinco vocales del español vistas en el espectrograma del programa *SpeechView* del CSLU/OGI (hablante femenino)

Hasta el momento, se ha hablado de las características de las vocales en general; sin embargo, existen diversos fenómenos que se producen cuando dos vocales se encuentran contiguas dentro del habla. “It is widely assumed that, where identical adjacent elements occur across word boundaries, one is eliminated in fast speech. Similarly, the adjacency of two non-identical vowels generally surfaces as a monosyllabic group. More precisely, four types of processes that can occur when two or more vowels are in contact between words are: hiatus (-V.V-), reduction (-VV-), diphthongisation (G[10]V, VG), deletion (V)” (Aguilar, 2003). Por ejemplo, en la palabra *teatro* un hiato sería [te.á.tro], una reducción [tea.tro], una diptongación [tja.tro] y una pérdida [ta.tro].

A continuación, abundaremos sobre tres fenómenos que suelen tener una gran ocurrencia en el español hablado en México cuando dos vocales se encuentra contiguas dentro de una misma palabra: diptongos, hiatos y diptongación de hiatos.

4.2. Diptongos e hiatos

4.2.1. Diptongos

Resulta necesario empezar este apartado definiendo lo que se entiende por diptongo. “Las vocales *i, u,* combinadas entre sí o acompañada cada una de ellas por otra vocal, dentro de una misma palabra, forman el grupo fonético que se llama diptongo. La *i* y la *u* se

[10] En la literatura filológica hispánica, se ha utilizado erróneamente el anglicismo *glide* para referirse al grupo formado por las semivocales y semiconsonantes del español. Un término más correcto es “paravocal”, que utiliza Cuétara (2004) a partir del artículo de Whitley (2000), “Las paravocales españolas, el hiato y la abertura de la conjunción”.

pronuncian, según queda dicho, como semivocales, cuando van al fin del diptongo, y como semiconsonantes, cuando van al principio" (Navarro Tomás,1918:65).

Por otro lado, Quilis (1981:178) opina que "en español, normativamente, se acostumbra a considerar como diptongo la unión en la misma sílaba de: 1. /i, u/ + /e, a, o/; 2. /e, a, o/ + /i, u/; 3. /i/ + /u/; 4. /u/ + /i/".

Por último, Salporta (1956) lo define de la siguiente manera. "The traditional statement regarding complex syllabic nuclei seems quite adequate: any (stressed or unstressed) vowel may be flanked by /i/ or /u/ to form a diphthong (except that /i(/ may be flanked only by /u/, and /i(/ may be flanked only by /i/".

Según estas definiciones, los diptongos siempre estarán conformados por la vocal palatal cerrada /i/ o la vocal velar cerrada /u/.

Además, Navarro Tomás, en su definición de diptongo, menciona que dependiendo del lugar en el que aparezcan las vocales /i, u/, serán consideradas como semivocales o semiconsonantes: Al respecto, Llisterri y sus colaboradores opinan que, "the vowels /i/ and /u/ have allophonic variants −known as semiconsonants or semivowels− according to their nuclear or peripheral position in the syllable" (Llisterri *et al.*,1993).

Por lo tanto, los sonidos [j], [i9], [w], [u9] (paravocales) son considerados, como se ha visto, alófonos de las vocales altas. Alarcos (1950:159) dice que los sonidos [j], [i9], [w], [u9] de los diptongos son, en general, simples variantes de los fonemas /i/ y /u/, respectivamente.

Las vocales débiles, al formar diptongos con vocales más abiertas, se convierten en paravocales. Al respecto, Gili Gaya (1975) dice que "en los diptongos, la vocal más abierta representa el punto vocálico de la sílaba; la más cerrada se halla en la tensión o en la

distensión. El hablante tiende a extremar la diferencia que entre ambas exista en su grado de abertura, bien abriendo más la abierta, bien cerrando más la cerrada, o ambas cosas a la vez. Tal es el caso, en español, de las vocales extremas *i, u,* llamadas *débiles* porque al formar diptongo con las más abiertas *a, e, o* (*fuertes*), éstas constituyen el núcleo silábico y aquéllas quedan en posición inicial (tensiva) o final (distensiva). En estas condiciones *i, u,* se abrevian, al mismo tiempo que estrechan su articulación hasta el punto de perder en parte su naturaleza vocálica y convertirse en semiconsonantes y semivocales: *j, w, i̯, u̯*" (117).

Hasta el momento, se ha examinado que las vocales altas que conforman un diptongo se clasifican en semiconsonantes o semivocales; a su vez, los diptongos se clasifican en crecientes y decrecientes. "Spanish has rising diphthongs, formed by the glides [j] or [w] plus a syllabic nucleus, and falling diphthongs, formed by a syllabic nucleus plus the glides i or [u̯]" (Martínez-Celdrán, 2003).

Los casos en que el diptongo está formado por /i, u/ + /e, a, o/ se llaman crecientes, ya que los órganos se desplazan hacia la abertura, y cuando el diptongo está formado por /e, a, o/ + /i, u/ se llaman decrecientes, ya que los órganos se desplazan hacia el cierre.

Quilis (1993:179) opina que, por un lado, en los diptongos crecientes la vocal alta (sea /i/ o /u/) se encuentra como margen silábico y recibe el nombre de semiconsonante, mientras la vocal que esta en segunda posición forma el núcleo silábico. La semiconsonante ocupa una posición silábica prenuclear y se transcribe fonéticamente como [j] (alófono en función silábica prenuclear de /i/) o como [w] (alófono en función silábica prenuclear de /u/). Por otro lado, en los diptongos decrecientes la vocal alta ocupa una posición silábica postnuclear y recibe el nombre de semivocal. Se transcribe fonéticamente como [i̯] (alófono en función postnuclear de /i/) o como [u̯] (alófono en función silábica postnuclear

de /u/). En estos casos, la vocal que forma el núcleo silábico está en primera posición y reúne las mejores condiciones fónicas de todos los segmentos vocálicos que forman la sílaba: tendrá una mayor abertura, mayor tensión, mayor intensidad, mayor duración y mayor perceptibilidad.

Se tiene así que, como se observa en el cuadro 2, existen ocho diptongos crecientes y seis decrecientes en el uso actual del español ya que, la combinación de dos vocales fuertes no se considera diptongo pero sí la combinación de las dos vocales débiles /i/ + /u/, /u/ + /i/, haciendo semiconsonante la que se encuentra en el primer elemento del diptongo.

Diptongos creciente	Diptongos decrecientes
ja	**aj**
je	**ej**
jo	**oj**
wa	**aw**
we	**ew**
wo	**ow**
ju	
wi	

Cuadro 3. Diptongos crecientes y decrecientes del español de México

En español, no es fácil determinar si una secuencia vocálica constituye o no un diptongo; a pesar de que se dice que si el acento está en las vocales fuertes se trata de un diptongo, se producen ciertas excepciones. Si se utiliza un espectrógrafo para conocer los formantes que se originan, observamos que "se produce un cambio lento de la transición entre los formantes de las dos vocales cuando forman un grupo tautosilábico; por el contrario, un cambio rápido refleja una secuencia heterosilábica, siendo tanto más acusada la percepción de hiato cuanto más rápido sea el cambio, ya que éste actúa como límite silábico" (Quilis,1981:179). (Ver imagen 10).

Ya se ha descrito el caso de los diptongos; sin embargo, en el español se presenta otro fenómeno que puede producir un diptongo: el de las sinalefas. “Al grupo de vocales formado por el enlace de las palabras y pronunciado en una sola sílaba se le da el nombre de sinalefa” (Navarro Tomás, 1918:69).

Si en un diálogo se encuentran contiguas una palabra que termine en vocal y otra que empiece con vocal se puede dar el caso de un diptongo. Para que ello suceda, dichas vocales deben formar alguno de los 14 diptongos que existen en la lengua y que vimos en el cuadro 2. Un ejemplo sería el siguiente, *todo un invierno*, donde las palabras *todo* y *un*, al ser emitidas en habla continua forman el diptongo [ow]: [tó ðou9 ni7m ðje7r no]. De esta manera, se observa que un diptongo no sólo se forma en una palabra, sino también se produce en la unión de dos palabras que reúnan las características mencionadas.

Asimismo, en el caso de una palabra que termine con vocal fuerte y otra que empiece con vocal fuerte se puede generar un hiato. Por lo tanto, cuando existen dos vocales contiguas dentro de una palabra, no sólo se genera un diptongo, sino que puede producirse un hiato. Observemos a continuación la diferencia que existe entre estos dos fenómenos.

4.2.2. Hiatos

Al igual que el apartado anterior, comenzaremos por definir el fenómeno; es decir, qué es un hiato. Navarro Tomás (1918:66) opina que con frecuencia aparecen juntas, dentro de una misma palabra dos vocales que no forman diptongo, sino que por tradición gramatical constituyen sílabas distintas. Define al efecto prosódico que produce la pronunciación de

las vocales colocadas en dicha posición como hiato. Por otro lado, Martínez Celdrán (1984:221) define la palabra hiato como dos vocales juntas en que ninguna de las dos ha perdido su autonomía y ambos son núcleos silábicos. Gili Gaya (1975:117) menciona que ocurre un hiato cuando dos o más vocales seguidas se pronuncian sin formar una sílaba única. Finalmente, Quilis hace una diferencia entre diptongo e hiato y dice que una "secuencia de dos o tres vocales puede estar comprendida en una sílaba o dividida en sílabas distintas. En el primer caso, constituye un *diptongo* (/ói/ *hoy*) o un *triptongo* (/bwéi9/ *buey)*; en el segundo, un hiato (/oí/ *oí*, /léo/ *leo*, /féa/ *fea*)" (1993:178).

Por lo tanto, la diferencia entre el hiato y el diptongo es que, mientras que en un diptongo las vocales se encuentran contiguas en la misma sílaba, en un hiato las vocales se encuentran en distinta sílaba y cada una es núcleo silábico. Así, lo primero que se necesita saber es si las vocales se pronuncian en la misma o en distinta sílaba. Sin embargo, esto no es tarea fácil pues, "aun en el caso de que cada vocal forme por sí misma una sola sílaba, el paso de una vocal a otra vocal inmediata se hace siempre en nuestra pronunciación gradualmente y sin interrupción de sonoridad" (Navarro Tomás, 1918:147).

Ante esta dificultad, comenzaremos por describir las características de un hiato. Éste se puede formar a partir de las 14 combinaciones de vocales que ocurren en un diptongo y, además, por las combinaciones que ocurren entre vocales fuertes. "Cuando la secuencia vocálica está formada por dos vocales medias /eo/, /oe/, o una media y otra baja, o viceversa, /ea/, /oa/, /ae/, /ao/, cada una de ellas es núcleo de una sílaba diferente, formando, por lo tanto, un hiato: *céreo, aseo, soez, beato, toalla,* etc" (Quilis, 1993:184). Sin embargo, Gili Gaya (1975:120) opina que en cuanto a las combinaciones que existen entre vocales fuertes sus diferencias relativamente pequeñas en el grado de abertura

motivan numerosas vacilaciones entre el hiato y el diptongo, que se suman, entre otras cosas, a las producidas por las variantes dialectales, por el acento, y por la mayor o menor rapidez y esmero de la dicción. Si se comparan las pronunciaciones de los ejemplos que Gili Gaya propone (1975) *al-co-hol* y *al-col, al-de-a-no* y *al-dea-no, pe-or* y *peor, re-al* y *real, bo-a-to* y *boa-to, le-er* y *ler*, podemos observar que se tratan de diptongos no consolidados, a los cuales "la gramática tradicional da el nombre de *impropios*, por no haber en ellos semiconsonante o semivocal" (Gili Gaya, 1975:121).

Este tipo de hiatos es justamente el que más se diptonga en el corpus que se analiza en este trabajo: suelen diptongarse por su rápida pronunciación y poco énfasis; los resultados se encuentran en el siguiente capítulo (§5.4).

Si recurrimos al espectrograma para observar este fenómeno, encontramos que cuando la transición de un formante a otro es lenta y su duración larga, se trata de un diptongo, mientras que si la transición es rápida y su duración breve, se trata de un hiato. En la imagen 10 se puede observar esta diferencia.

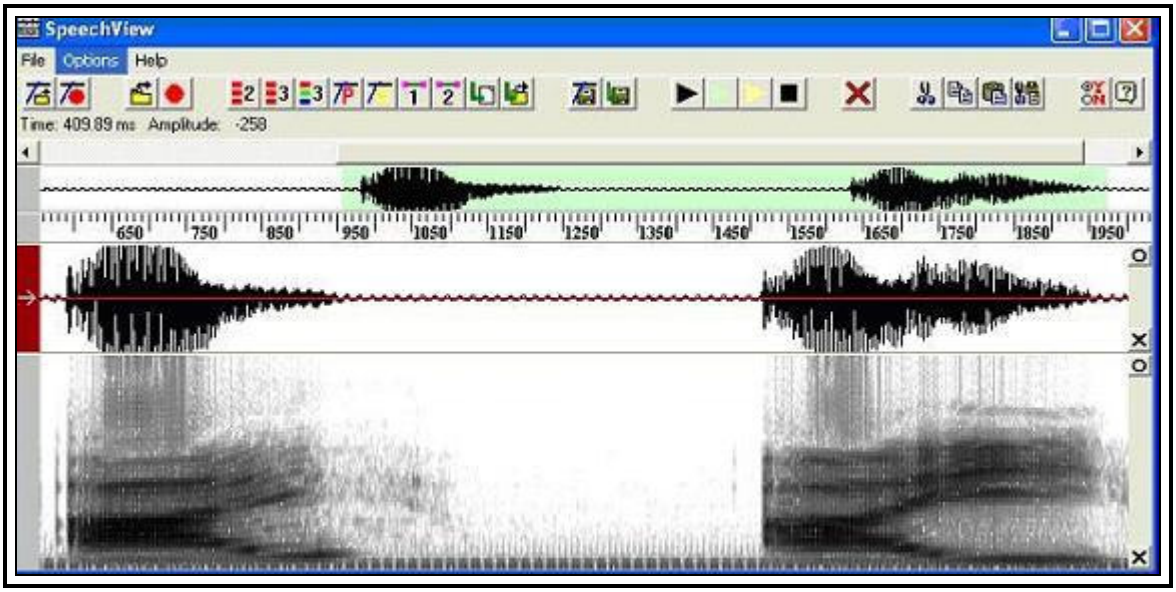

Imagen 10. Las palabras *hay* (diptongo) y *ahí* (hiato) vistas en un espectrograma

Debido a que en un hiato ambas vocales son núcleo silábico, conservan su autonomía y, por lo tanto, "espectrográficamente, esas dos vocales mantienen sus propios formantes bien diferenciados; se saltará de un formante a otro bruscamente. En cambio, en el diptongo, la glide ha perdido su propia autonomía, se ha sometido a la vocal que forma el núcleo silábico y esa dependencia se muestra espectrográficamente como si el formante de la glide fuese un mero apéndice del núcleo silábico; es, en definitiva, una transición muy prolongada del segundo formante de la vocal de la que depende" (Martínez-Celdrán,1984:221). Por lo tanto, y retomando lo que ya se ha dicho, la diferencia entre un diptongo y un hiato es la posición de las vocales en la sílaba. Si las vocales se encuentran contiguas en la misma sílaba se trata de un diptongo, mientras que si las vocales se encuentran contiguas en distinta sílaba se trata de un hiato. Además, espectrograficamente esta diferencia se observa en la transición de los formantes vocálicos.

En cuanto a la frecuencia de este fenómeno, Navarro Tomás (1918:159) opina que en el español la analogía favorece el hiato, más aún en las formas verbales, cuando dentro

del mismo verbo de que se trata hay casos en que las vocales *i, u,* llevan el acento fuerte y pone como ejemplos: *fiar, fianza (fían)*; *guiaba (guía)*; *liamos (lías)*. Sin embargo, en español existe una tendencia a reducir los hiatos a diptongos. El lenguaje lento, el acento enfático y la posición final favorecen el hiato, mientras que una pronunciación en tono corriente y familiar beneficia al diptongo. Ya que el habla suele producirse de la segunda manera, dos vocales contiguas serán susceptibles de reducirse a una sola sílaba y, por lo tanto, a formar un diptongo. Este es el fenómeno que se conoce como diptongación de hiatos, y suele ser muy frecuente en el español -especialmente de México. Es así como se explican ciertas pronunciaciones como [tjátro], *teatro*, [pio**7**r], *pjor*, [trái**9**], *trae*, etc. Gili Gaya explica esta situación y dice que "cuando se forma el diptongo con vocales fuertes, la más abierta tiene intensidad y duración normales pero la más cerrada se debilita y abrevia, al mismo tiempo que tiende a cerrar más su articulación" (1975:120).

4.2.3. Diptongación de hiatos

Matluck (1951:17) hace un estudio de las características del español del Valle de México y menciona que, además del seseo[11] (general en toda Hispanoamérica) y del yeísmo (casi general en toda Hispanoamérica), el habla popular del Valle se caracteriza por siete fenómenos, entre los que figura la diptongación de hiatos.

Por lo tanto, la diptongación de hiatos se encuentra entre los fenómenos característicos del habla popular del Valle de México. "Quien revise con atención la bibliografía sobre el español mexicano notará que el fenómeno de diptongación de hiatos,

[11] Como es bien sabido, el seseo se refiere a la no distinción entre la consonante alveolar fricativa sorda y la consonante interdental fricativa sorda castellana, haciendo general la primera. Por otro lado el yeísmo se refiere a la no distinción entre la consonante palatal fricativa sonora y la lateral palatal castellana.

por una parte, se menciona desde los primeros estudios conocidos y, por otra, que frecuentemente se atribuye a la pronunciación del español de todos los mexicanos y no sólo de alguno de sus dialectos" (Moreno,1994). Si se presta atención en el habla cotidiana, es notable que los hablantes del español tienen una preferencia a diptongar los hiatos, en casos como el grupo inacentuado de [ae], en donde la /e/ cambia a /i/ y se rompe constantemente el hiato por ejemplo, [kai**9**rá], *caerá*, [kai9rémo7s], *caeremos*, [trai**9**rán], *traeran*. Las causas de este fenómeno las explica Quilis (1993) en su *Tratado de fonología y fonética española*, donde dice que "esta tendencia antihiática del español responde a dos causas que se complementan: 1) una se refiere al límite silábico: la secuencia silábica ideal es CV-CV-CV, donde una consonante (C), que es más cerrada, normalmente, que una vocal, marca la frontera o límite silábico. En la secuencia heterosilábica V-V', el límite silábico está muy débilmente señalado: sólo por la transición formántica, más o menos rápida que hay entre sus vocales; para evitar ambigüedades, la lengua se vale de cualquiera de los dos medios antes indicados: a) suprimir el límite silábico, convirtiendo el hiato en diptongo; b) reforzar el límite silábico, introduciendo en él una consonante. 2) La otra causa se debe a un principio de economía, que, de un modo u otro, suele estar presente en los cambios fonéticos: en este caso, se trata del gasto de aire, que, como se ha visto, es mayor en las vocales altas y en las acentuadas: para pronunciar [aí] *ahí*, se necesita mucho más aire que para [a i9], pronunciación frecuente de *ahí*" (190).

Finalmente, a continuación se confrontan dos opiniones opuestas; una que considera a los diptongos como monofonemáticos; la otra, como bifonemáticos.

4.3. Dos visiones: Trubetzkoy y Alarcos

Como ya se mencionó en la introducción de este trabajo, el propósito principal de esta tesis es el de proponer la etiquetación de los diptongos como una sola unidad y no como dos, como hasta ahora se hecho; por tanto, es importante contrastar las ideas que tienen dos autores importantes, Trubetzkoy y Alarcos, de considerar a los diptongos como monofonemáticos o bifonemáticos respectivamente.

Por un lado, Trubetzkoy (1964:49) opina que, "sólo pueden ser interpretados como monofonemáticos los grupos de sonidos cuyos componentes, en la lengua considerada, no se reparten en dos sílabas, son producidos por un único movimiento articulatorio y cuya duración no excede la duración normal de los sonidos simples". Según esta definición, el grupo de sonidos que componen un diptongo debe ser considerado como monofonemático. Sin embargo, Alarcos (1950:152) dice que los componentes [a], [a6], [e7], [e], [o], [o7], [u], [i] de los diptongos españoles son realizaciones diversas de los fonemas vocálicos, pues no hay conmutación entre [a] y [a6], entre [e7] y [e], etc. Por lo tanto, según este autor, si el grupo de sonidos que componen un diptongo son considerados únicamente como alófonos de las vocales, entonces carecen de valor monofonemático y sólo son combinaciones de los cinco fonemas vocales con otro elemento.

Por otra parte, Perissinotto (1975:33) coincide con Alarcos, pues también opina que los diptongos deben considerarse como bifonemáticos, ya que las semivocales y las semiconsonantes [j], [i9], [w], [u9] son variantes de /i/ e /u/.

De esta manera, se contrastan las opiniones de cada autor de considerar a los diptongos monofonemáticos o bifonemáticos y, a pesar de que son definiciones distintas,

existe una idea de uno y otro en la que coincidimos en este trabajo. Por un lado, es cierto que los diptongos se componen de realizaciones diversas de fonemas vocálicos, y por otro lado, también es cierto que son producidos por un único movimiento articulatorio.

No obstante, coincidimos más con la idea de Trubetzkoy de considerarlos como monofonemáticos pues, a pesar de que [j], [i9], [w], [u9] son alófonos de las vocales /i/ e /u/, la transición de sus formantes se da de manera lenta. Esto último, es consecuencia de la proximidad con que se produce uno y otro sonido y, por lo tanto, la paravocal pierde su autonomía y se somete a la vocal que forma el núcleo silábico. El formante del núcleo silábico se vuelve una continuación de la paravocal, en el caso de los diptongos crecientes, y a la inversa en el caso de los diptongos decrecientes.

De tal suerte, los diptongos deberían considerarse monofonemáticos, ya que no son puramente dos fonemas distintos, sino que al estar en la misma sílaba pierden su autonomía. Además, Navarro Tomás (*apud.* Gili Gaya) coincide con Trubetzkoy, ya que en sus *Estudios de Fonología española* de 1946, estima que "los diptongos son unidades fonológicas capaces de crear oposiciones como *peina-pena-pina, tuerca-terca-turca, uso-hueso-eso,* etc. Siendo así, las semivocales y las semiconsonantes, que sólo aparecen en diptongos y triptongos, no pueden ser consideradas por sí solas como fonemas, sino como componentes de un fonema compuesto" (1975:120).

Consecuentemente, no queda más que decir que el caso de los hiatos no podría ser un grupo de sonidos monofonemático, ya que la transición de sus formantes es de manera rápida: cada vocal es núcleo silábico y conserva su autonomía. Asimismo, los hiatos no son capaces de crear oposiciones, como en el caso de los diptongos.

Es por esto que, en la etiquetación del Corpus DIME que se propone en este trabajo, las etiquetas de diptongos, de sinalefas que formen diptongos y de diptongación de hiatos se transcribirían como monofonemáticos; es decir, utilizando una sola etiqueta para las dos vocales que conforman el diptongo, y no dos como hasta ahora se ha hecho. Para esto se tendrían que incluir 14 modelos acústicos más en el reconocedor de habla. En el caso de los hiatos, esto no sería viable por lo que se explicó anteriormente. Y en el caso de las sinalefas en el que se formen diptongos, es necesario, como ya se mencionó (§2.), crear un diccionario en el que el reconocedor detecte este fenómeno como diptongo y no como dos palabras aisladas, así como incluir la información necesaria en los modelos del lenguaje.

Para recapitular, en esta sección se hizo una revisión de las vocales, dando sus rasgos principales y su frecuencia de aparición según la teoría filológica tradicional. Además, se revisó cada uno de los fenómenos en los que aparecen dos vocales contiguas: diptongos, sinalefas, hiatos y diptongación de hiatos. Por último, se explicó con la teoría de Trubetzkoy cómo se pueden considerar los diptongos monofonemáticos y, de esta manera, presentarlos en el etiquetado fonético computacional como una unidad en un corpus que tendrá como finalidad la obtención de un reconocedor de habla.

5. ANÁLISIS DE DATOS

Debido a que el interés principal de este trabajo de tesis es el estudio de los diptongos del español de México, en este capítulo se hará el análisis de cinco diálogos del Corpus DIME, con el objetivo de obtener datos acerca de la presencia de diptongos en el habla del español de México, y compararlos con la teoría clásica que se desarrolló con mayor precisión en el capítulo anterior.

Se dará una introducción en donde se hablará del Proyecto DIME y el corpus con el cual se trabajó, para posteriormente exponer el análisis de datos y las conclusiones a las cuales se llegó.

5.1. El proyecto DIME

Actualmente, en el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS) de la Universidad Nacional Autónoma de México, se lleva a cabo el proyecto "Dialogos Inteligentes Multimodales del Español" (DIME), en el que se trabajan tres módulos, "el primero incluye la recopilación, transcripción y etiquetación de un corpus multimodal[12] en un dominio de diseño; el segundo consiste en la construcción de un sistema de reconocimiento del habla para el español especializado en dicho dominio, y el tercero en la definición de una gramática del español que contemple los fenómenos

[12] Multimodal: que incluye video, grabaciones, sonidos, imágenes; es decir, multimedia.

gramaticales y el lexicón observados en el corpus, así como la implementación de su proceso de análisis gramatical" (Pineda *et al*., 2001). El objetivo principal, a largo plazo, es el desarrollo de sistemas conversacionales en español.

El trabajo interdisciplinario entre lingüistas y computólogos en este proyecto es fundamental. Por un lado, el lingüista aporta la información fonética, fonológica, prosódica o gramatical, según sea el caso. Por ejemplo, una de las herramientas básicas es el alfabeto fonético computacional; el lingüista diseña y perfecciona éste, incluyendo las realizaciones necesarias de cada fonema. Por otra parte, se transcribe el corpus en formato electrónico (transcripción fonética computacional) utilizando la herramienta del *SpeechViewer*, del (CSLU), que permite observar la onda sonora y su respectivo espectrograma con el fin de etiquetar cada uno de los alófonos que se generan a lo largo del discurso.

Por otro lado, el computólogo, utiliza las etiquetas del corpus para entrenar al sistema y crea los modelos acústicos con las herramientas necesarias (entre los sistemas más reconocidos y funcionales están HTK y Sphinx) (Pineda *et al*. 2001), con el fin de producir un reconocedor de habla.

Pineda *et al*. (2001) señalan que "dentro del contexto del proyecto se desarrolla un módulo de reconocimiento del habla para el español de México en el dominio de aplicación. El objetivo de este módulo es desarrollar los modelos acústicos-fonéticos, los diccionarios de pronunciación y los modelos del lenguaje, para lo cual se utiliza el *CSLU Toolkit* y el sistema *HTK* basado en modelos ocultos de Markov. Con el propósito de crear los diccionarios de pronunciación se definió un alfabeto fonético *Mexbet*".

Para desarrollar los modelos acústicos-fonéticos, los diccionarios de pronunciación y los modelos del lenguaje, se empieza con la creación de un corpus, que en este caso se llevó a cabo en el dominio de cocinas. Una cosa importante que se requiere para la creación

de éste es el uso del lenguaje natural; "there are several important considerations that have to be taken into account for the construction of conversational working prototypes. First, the application domain should be complex enough to merit the use of natural language but, at the same time, as simple as possible to be able to model it with current computational technology" (Villaseñor *et al.*, 2001).

El Corpus DIME es, por tanto, lo suficientemente complejo para que el uso del lenguaje natural sea útil, pero al mismo tiempo lo suficientemente simple para poder modelar el procesamiento del lenguaje.

Posteriormente, se hace el etiquetado fonético del corpus, alineando temporalmente la señal sonora y las etiquetas. Para ello, se utiliza un alfabeto fonético computacional en código ASCII; en el caso del español de México, se ha utilizado, como vimos (§3.2.1.), Mexbet (Uraga y Pineda, 2000; Cuétara, 2004). Además de las etiquetas de alófonos, el corpus se etiqueta a nivel ortográfico y fonético. Un trabajo futuro buscará el etiquetado del corpus a nivel prosódico, con el fin de crear modelos para la información entonativa y emocional del discurso, pues "en una conversación a través de diálogos, se transmite naturalmente información entonacional y emocional, ya que el habla es rica en técnicas interactivas que garantizan que el receptor entiende lo que se expresa, incluyendo las expresiones formales y gestos físicos y vocales" (Olivier, 1999:16).

Simultáneamente al trabajo de etiquetado, se crean los diccionarios de pronunciación, que contemplan el lexicón del corpus, así como las posibles combinaciones que existen entre los componentes de éste. Los diccionarios de pronunciación son los listados de las palabras que aparecen en el dominio especificado en cada proyecto (un ejemplo en el Corpus DIME son las palabras *fregadero, cocina, lavatrastes*, etc.). Las listas se elaboran con las pronunciaciones que puede recibir cada una de esta palabras en la

lengua hablada. Por ejemplo, la vocal /o/ final o las consonantes /g/ y /d/ de la palabra *fregadero*, podrían sufrir debilitamientos; de tal suerte, en el diccionario de pronunciación se incluyen la forma prototípica y las variaciones que pudiera presentar esta palabra en el habla.

fregadero	**[f r(e G a D e r(o]**
	[f r(e G a D e r(O]
	[f r(E G a D e r(O]
	[f r(E G a D e r(o]
	[f r(E G a d_c d e r(o]
	...
	...
	...

Cuadro 4. Muestra del diccionario de pronunciación de un reconocedor de habla: transcripción en Mexbet de la palabra *fregadero*

Finalmente, en el proceso de formalización computacional se entrena con todas estas herramientas (etiquetas, diccionario, etc.) al sistema, para crear los modelos acústicos. Una vez que se cuenta con suficientes muestras de cada alófono, el sistema es capaz de modelarlos primero, para identificar después cada uno de ellos y llevar a cabo el reconocimiento de habla. Así, resalta la importancia de que el corpus tenga un equilibrio de fonemas; es decir, que todas las formas del sistema fonológico, y todas las realizaciones fonéticas concretas, aparezcan en una cantidad suficiente y proporcional.

Además de los modelos acústicos y del diccionario, el reconocimiento de habla consta de los modelos del lenguaje que, como ya se mencionó (§2.), contienen información sobre la posible combinación entre las palabras del corpus.

Por último, en el proceso del modelado computacional del lenguaje se construye un programa computacional para auxiliar en tareas de diseño que permita sostener una

conversación con el sistema en el dominio de aplicación. Debido a que parte del trabajo de un lingüista en la creación de un reconocedor de habla es la obtención del corpus, a continuación se describen ciertas características que se deben tomar en cuenta, así como el proceso que se llevó a cabo para la creación del Corpus DIME.

5.2. El Corpus DIME

Es necesario comenzar este apartado señalando lo que se entiende por corpus lingüístico. Llisterri (1991) menciona que "suele llamarse corpus al conjunto de realizaciones sonoras de la lengua que serán objeto de estudio", y posteriormente amplía esta definición diciendo que "tanto si es oral como escrito, un corpus puede concebirse como un conjunto estructurado de materiales lingüísticos en el que se distinguen diversos niveles de representación correspondientes a diferentes grados de elaboración de los datos que lo constituyen".

En un principio, el Corpus DIME no tuvo en esencia el propósito de crear modelos acústicos, sino que su objetivo principal era estudiar la interacción humano-computadora, ver qué léxico se desprendía para el dominio de cocinas y modelar la gramática del español, entre otras cosas (Pineda *et al*., 2001). Es por esto que, para la elección de hablantes, no se llevó a cabo ningún criterio lingüístico específico, hecho que puede resultar un tanto desventajoso para el trabajo del corpus, ya que los grupos de informantes no son homogéneos y, por lo tanto, no constituyen globalmente una variable controlada. No se determinó ni la edad, ni el nivel socioeconómico, ni la procedencia de cada uno de ellos.

No obstante, podemos tomar a nuestro favor la idea que de Lope Blanch (1969) cuando dice que "las grandes ciudades de nuestro tiempo, por su inmenso poder de

atracción, reúnen en un seno a hablantes de muy diversa procedencia regional, y actúan así como crisol en que se funden las distintas tendencias o peculiaridades idiomáticas de todo el país". Así, se toma la desventaja de la ausencia de diseño de hablantes en el Corpus DIME como una ventaja, en el sentido de que nos arroja una muestra aleatoria del habla de la ciudad de México.

Una de las desventajas que no ha sido fácil subsanar en el Corpus DIME, es la ausencia de un diseño alofónico previo que arroje la mayor variedad posible de realizaciones para que puedan crearse los modelos acústicos. De tal suerte, si los datos han de ser contrastados o ampliados, se tiene que recurrir a un corpus alternativo.

Actualmente existe el Corpus DIMEx100 (Pineda *et al.* 2004), realizado dentro del proyecto DIME, en el cual sí se tomó en cuenta la mayor cantidad posible de alófonos del español de México. Se grabó a 100 hablantes, los cuales a su vez grabaron 50 oraciones distintas y 10 oraciones comunes a todos; la distribución fonética y fonológica fue cuidada.

En el caso del corpus DIME, éste se realizó de manera oral y dentro del ámbito del habla espontánea, al igual que el corpus que utiliza Lope Blanch en el *Estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y la Península Ibérica*, donde se hizo el acopio de información sirviéndose de materiales grabados. "En las grabaciones se recogerán, fundamentalmente, conversaciones libres entre dos informantes" (Lope Blanch, 1969). De la misma manera, el corpus DIME recoge información del habla que se obtiene entre el sistema y el usuario, ya que se lleva a cabo a partir de conversaciones libres; esto proporciona una información entonacional y emocional, a diferencia de un corpus en el que únicamente se leen las oraciones. De tal suerte, el Corpus DIME arroja datos, además de fonéticos, prosódicos, y de la estructura de los diálogos orientados a la solución de problemas.

Como se mencionó (§3.3.1.), con el fin de considerar a los hablantes que participan en el Corpus DIME como posibles usuarios en un sistema de diálogo, el protocolo que se llevó a cabo fue el del "Mago de Oz". "El método del Mago de Oz se suele emplear en la creación y evaluación de prototipos con los siguientes objetivos:

1. **Probar la eficacia del sistema [...]**
2. **Analizar el comportamiento de las personas en la interacción con una máquina para extraer la información lingüística que necesitará el sistema real cuando haya de interpretar los mensajes del usuario [...]**
3. **Entrenar el módulo de reconocimiento, ya que en la mayoría de las aplicaciones se requiere el reconocimiento del habla espontánea. El reconocedor puede mejorar si se incorpora información sobre ciertos fenómenos lingüísticos propios de este estilo de habla: repeticiones, inserciones que truncan el discurso, relajaciones en la articulación de los sonidos, alargamientos de vocales, ruidos del usuario, vocalizaciones, etc." (Llisterri, 2003).**

Además, no debe hacerse de lado la importancia que tiene el papel que juega cada uno de los elementos en un discurso: el hablante, el mensaje y el receptor, "three elements are distinguished: the message -which in fact corresponds to the corpus obtained-, the speaker and the listener and the relationship they maintain, and the context and the channel, that in the situations to which the classification is restricted correspond to the source from which the corpus has been obtained" (Aguilar, 1994).

Así, la tarea que se realizó, como ya se mencionó (§3.3.1.), fue de la siguiente manera: el "Mago" se encontraba en un cuarto aislado mientras que el sujeto se hallaba en otro cuarto. Al igual que el hablante; ambos observaban la misma imagen en sus pantallas. El objetivo del hablante era crear una cocina idéntica a la que estaba en una imagen que se le había proporcionado anteriormente. Villaseñor *et al.* (2001) opinan que una de las razones para escoger la tarea del dominio de cocinas frente a otras tareas radica en que es mucho más común para la gente, conocer cómo se ve una cocina.

Para que el "Mago" haga el papel del sistema lo mejor posible, recibe una lista de lineamientos que debe llevar a cabo a la hora del experimento. Algunas de ellas son: "The

system cannot interrupt the user when he or she is speaking, the system cannot infer more than what is obviously said, the system cannot understand too long, complicated and unclear sentences" (Villaseñor *et al*., 2001).

Tomando en cuenta los lineamientos que se han mencionado, se llevó a cabo el Corpus DIME. En él, se hicieron 15 experimentos con 15 sujetos diferentes, y algunos experimentos complementarios; se obtuvieron 31 diálogos útiles, con un total de 27, 459 instancias de palabras (886 en promedio por diálogo), 5779 enunciados (185 en promedio por diálogo), 3606 turnos (115 en promedio por diálogo) y 7:10 Hrs. (14 minutos en promedio por diálogo)" (Pineda *et al*., 2001).

El diálogo se grabó utilizando el programa *Wave Studio* de la compañía *Creative*, y posteriormente se llevó a cabo la segmentación del corpus en unidades mínimas que se les conoce como elocuciones. Éstas pueden responder a segmentos verbales o no verbales entre los dos interlocutores. El material que se consiguió tiene la característica de estar en lenguaje natural mediante diálogos entre un sistema y un usuario. Se toman en cuenta fenómenos del habla espontánea, como interjecciones, silencios muy largos y repeticiones, entre otros.

Con el propósito de tener una idea sobre lo que es el Corpus DIME, en seguida se presenta un fragmento de uno de los diálogos:

Mago: ¿Quieres que desplace o traiga algún objeto a la cocina?
Sujeto: Ehhhhhh.
Sujeto: Sí.
Sujeto: Vamos a mover... este mueble... a la esquina del fondo... pero rotándolo de una vez para que queden las puertas de frente... ¿okey?
Mago: Okey.
Mago: ¿Quieres que mueva este objeto en esta esquina?
Sujeto: Ahá.
Mago: Okey.
Mago: ¿Así está bien?
Sujeto: Así está bien.

Una vez que se ha revisado en que consiste el Corpus DIME, cómo se construyó y cuáles son sus características, a continuación se muestran los datos que se obtuvieron con respecto a la frecuencia de diptongos en el español de México, así como el análisis de ellos.

5.3. Análisis de datos

Los datos registrados en este trabajo fueron obtenidos de cinco diálogos del Corpus DIME: 2, 4, 5, 6 y 8. Se eligieron estos cinco diálogos de los 31 que conforman el total del Corpus DIME, ya que son los que se encuentran etiquetados manualmente y pueden facilitar el trabajo de etiquetado de los diptongos como una unidad. En cuanto al sexo de los informantes, encontramos que son dos mujeres y tres hombres. Además, el "Mago" pertenece al segundo grupo; por tanto, existen en los diálogos analizados, 2 mujeres y 4 hombres.

Ya que el interés principal de esta tesis es el estudio de los diptongos del Corpus DIME, se realizó un conteo de todas las situaciones en las que aparecieran dos vocales contiguas, ya sea al final e inicio de una palabra (sinalefa), dentro de una palabra en la misma sílaba (diptongo), dentro de una palabra pero en distinta sílaba (hiato) y, finalmente dentro de una palabra en una misma sílaba cuando la norma la separa en dos sílabas (diptongación de hiatos).

	Diálogo 2
Sinalefas	**101/230**
Porcentaje	**43.91 %**
Diptongos	**112/230**
Porcentaje	**48.69 %**
Hiatos	**5/230**
Porcentaje	**2.17 %**
Diptongación de hiatos	**12/230**
Porcentaje	**5.21 %**

Cuadro 5. Contabilización de fenómenos relacionados con la contigüidad de vocales en el diálogo 2 del Corpus DIME

En el cuadro 4 se observa cómo en un sólo diálogo aparecen 230 casos en los que existen dos vocales contiguas, siendo los diptongos los de mayor frecuencia, pero con sólo aproximadamente el 4 % más que las sinalefas. Son los hiatos y la diptongación de hiatos los que aparecen en un porcentaje mínimo en el diálogo 2. En el cuadro 5 se concentran las cifras de los diálogos restantes.

	Diálogo 4	Diálogo 5	Diálogo 6	Diálogo 8
Sinalefas	**63/158**	**56/158**	**71/185**	**50/133**
Porcentaje	**39.87 %**	**35.44 %**	**38.78 %**	**37.59 %**
Diptongos	**75/158**	**87/158**	**107/185**	**68/133**
Porcentaje	**47.46 %**	**55.06 %**	**57.83 %**	**51.12 %**
Hiatos	**7/158**	**5/158**	**1/185**	**4/133**
Porcentaje	**4.43 %**	**3.16 %**	**1/185**	**3.00 %**
Diptongación de hiatos	**13/158**	**10/158**	**6/185**	**9/133**
Porcentaje	**8.22 %**	**6.32 %**	**3.24 %**	**6.76 %**

Cuadro 6. Contabilización de fenómenos relacionados con la contigüidad de vocales en los diálogos 4, 5, 6 y 8 del Corpus DIME

En los datos se observa que los diptongos son los que más número de veces aparecen en todos los diálogos, siendo en la mayoría, más de la mitad de todos los casos. Nuevamente las sinalefas son las que se acercan más a la cantidad de diptongos, con una

diferencia máxima del 20 %. Los hiatos y diptongación de hiatos continúan siendo los casos menos comunes.

En resumen, el resultado final que se obtiene de aparición de vocales contiguas en 5 de los 31 diálogos del Corpus DIME es el que aparece en la imagen 11:

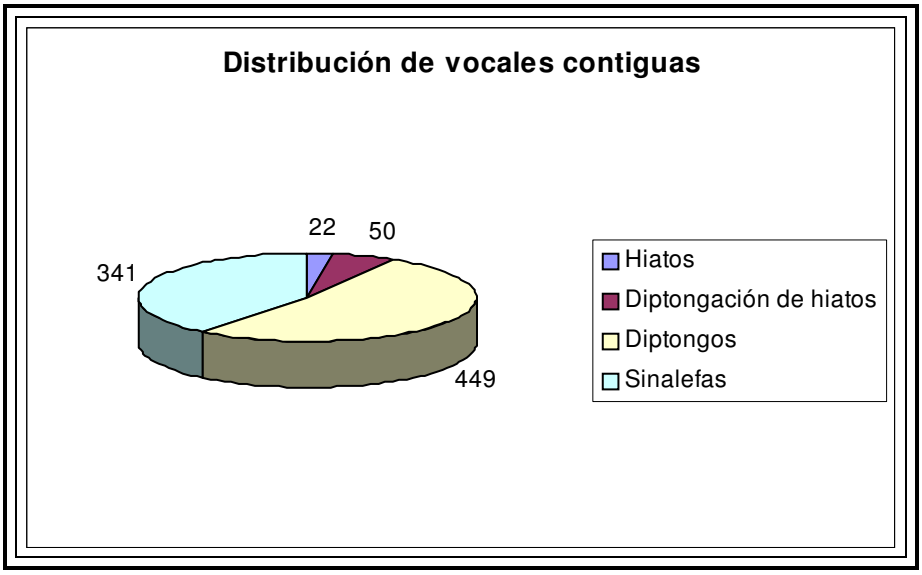

Imagen 11. Distribución de vocales contiguas en cinco de los 31 diálogos del Corpus DIME

En seguida se presenta el análisis de la frecuencia de combinación de vocales que se encuentran en los 449 diptongos que se obtuvieron en cinco de los 31 diálogos del Corpus DIME.

Diptongo	Cantidad	Porcentaje
je	**172**	**38.30 %**
we	**98**	**21.82 %**
ej	**91**	**20.26 %**
ja	**31**	**6.90 %**
aj	**19**	**4.23 %**
jo	**17**	**3.78 %**
wa	**16**	**3.56 %**
wi	**4**	**0.89 %**
oj	**1**	**0.22 %**

Cuadro 7. Cifras que representan las combinaciones de vocales en diptongo en el Corpus DIME

El cuadro 6 muestra que únicamente aparecen 9 de los 14 diptongos que existen en el español, siendo [iw], [aw], [ew], [ow], [wo] los cinco diptongos que no están presentes. Además, se observa que los cuatro diptongos más frecuente son diptongos crecientes, con excepción de [ej] que se encuentra en una posición decreciente. Se observa también que la vocal anterior media palatal es la que aparece en los tres primeros diptongos y, comparando estos resultados con el cuadro 7 que se muestra a continuación, se puede subrayar que la /e/ y la /a/ son las vocales con mayor frecuencia de aparición en el español.[13] Por lo tanto, los datos obtenidos en el Corpus DIME comprueban los porcentajes reflejados por 6 autores.

Vocal	Alarcos Llorach (1965)	Navarro Tomás (1946)	Zipf y Rogers (1939)	Guirao J. y García (1993)	Pérez (2003)	Cuétara (2004)	Promedio
/i/	**8.60**	**4.76**	**4.20**	**6.59**	**7.46**	**7.18**	**6.46**
/e/	**12.60**	**11.75**	**12.20**	**14.99**	**14.13**	**15.35**	**13.50**
/a/	**13.70**	**13.00**	**14.06**	**13.27**	**12.31**	**14.05**	**13.39**
/o/	**10.30**	**8.90**	**9.32**	**10.75**	**9.28**	**9.89**	**9.74**
/u/	**2.10**	**1.92**	**1.76**	**2.79**	**3.05**	**3.22**	**2.47**

Cuadro 8. Comparación de la distribución porcentual de la frecuencia de los fonemas del español

[13] Por el interés de este trabajo, únicamente se muestra la tabla de frecuencia de vocales; sin embargo, las tablas de todas las consonantes se encuentran en Pérez. 2003 (Ms.). "Frecuencia de fonemas", Universidad de Concepción, y en Cuétara. (2004), Pineda *et al*., (2004).

Aunque algunos de estos autores difieren en los porcentajes de dichas vocales, se puede sacar un promedio en donde se encuentra que la vocal media anterior es la más frecuente en el español en los estudios orientados al español de Latinoamérica, y la baja para el caso de los estudios orientados hacia el español peninsular. Además, existen estudios realizados por Quilis (1981), Llisterri y Mariño (1993) y Cuétara (2004) que constatan lo antes dicho.

Haciendo una comparación entre la frecuencia de vocales y de consonantes en el español, se puede decir que "vocales y consonantes se distribuyen en el sistema en proporciones casi similares, manifestando un porcentaje levemente más alto las consonantes, pero reducido si consideramos la cantidad de elementos que compone cada conjunto" (Pérez, 2003).

Se observa, además, que la cantidad de diptongos crecientes es mayor a la de diptongos decrecientes, obteniendo los siguientes resultados:

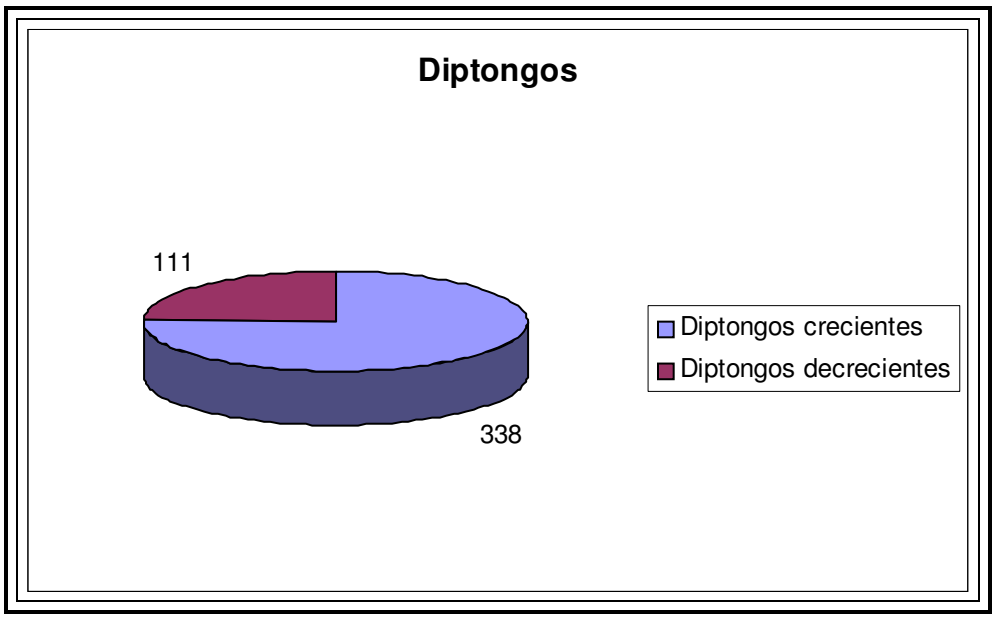

Imagen 12. Distribución de diptongos crecientes y decrecientes en cinco de los 31 diálogos del Corpus DIME

En cuanto a su posición en la sílaba, como se observa en el cuadro 8, se encontró que el diptongo [je] es el más frecuente en sílaba cerrada, es decir, un diptongo seguido de una consonante en la mima sílaba, por ejemplo /sjempre/; mientras que el diptongo decreciente [ej] es el más frecuente en sílaba abierta, por ejemplo /pejne] . Sin embargo, tanto el diptongo [we] como el [je] también se encuentran con una alta frecuencia en sílaba abierta. En el mismo cuadro, se observa, por el contrario, que los diptongos [ej], [aj], [wi] y [oj] son nulos en sílaba cerrada.

Diptongo	Abierta	Cerrada
je	**66**	**106**
we	**79**	**19**
ej	**91**	**0**
ja	**25**	**6**
aj	**19**	**0**
jo	**9**	**8**
wa	**3**	**13**
wi	**4**	**0**
oj	**1**	**0**

Cuadro 9. Distribución de diptongos en cinco de los 31 diálogos del Corpus DIME, según su posición en la sílaba

Por último, en cuanto a si el diptongo es acentuado o no acentuado, se observa en el cuadro 9 que, de los acentuados, el diptongo mas frecuente es [je], seguido de [we] con una gran diferencia sobre el resto de los diptongos acentuados. Por otro lado, [ej] representa el diptongo inacentuado con mayor número de apariciones y con una diferencia muy grande con respecto al resto de los diptongos inacentuados.

Por el contrario, tanto [ej] como [wi] no aparecen dentro de los diptongos acentuados, así como dentro de los inacentuados la frecuencia de aparición de [we], [aj], [wa] y [oj] es nula.

Diptongo	Acentuado	No acentuado
je	**170**	**2**
we	**98**	**0**
ej	**0**	**91**
ja	**2**	**29**
aj	**19**	**0**
jo	**9**	**8**
wa	**16**	**0**
wi	**0**	**4**
oj	**1**	**0**

Cuadro 10. Distribución de diptongos acentuados y no acentuados en cinco de los 31 diálogos del Corpus DIME

Hay un factor importante que se debe resaltar de inmediato: en nuestro caso, la alta frecuencia de aparición del diptongo [ej] se debe a una razón específica, ya que aparece las 91 veces en la palabra *okey*, siendo un extranjerismo adoptado con muchísima frecuencia en la lengua castellana, que sustituye a las afirmaciones *sí*, *estoy de acuerdo*, o *está bien*, según sea el caso. En este trabajo se encuentra en gran cantidad porque el "Mago" continuamente está asegurándose de que lo que realiza el sistema sea correcto para el usuario, y ambos utilizan esta palabra para afirmar que lo que se ha hecho es lo deseado. Además, ya que la aparición de este diptongo se debe únicamente a la palabra *okey*, los datos reflejan que aparece únicamente en sílaba abierta y es acentuado.

En lo que respecta al diptongo [je], que es el más frecuente dentro del corpus, Matluck en su estudio de *La pronunciación en el español del valle de México* encuentra que "lo normal en el Valle es diptongar la *e* acentuada y no deshacer el diptongo; es decir: *é* > *ié* (casi siempre por el vulgo, con bastante penetración en el habla semiculta), pero *ié* no da *é*, como en muchas regiones. Registramos las palabras *diferiencia, inociencia,* pero ningún ejemplo de *cencia, pacencia, etc*." (Matluck, 1951:33).

En el Corpus DIME no se encontró ningún dato como éstos; las realizaciones del diptongo [je] se dieron en palabras como *quieres*, *izquierda* y *superficie*, entre otras, que por ser parte de un corpus en el diseño de cocinas, se refieren principalmente a la localización de objetos en un espacio y, por lo tanto, suelen ser muy frecuentes.

Se debe tomar en cuenta también que los estudios realizados por Matluck fueron dentro de un habla de corte muy popular.

El caso de [we], que es el diptongo que aparece más frecuentemente después de [je], se debe a las mismas razones, pues las palabras en las que se presenta son *mueves*,

mueble(*s*), *puerta*, *puedas* y *muestras*, entre otras, lo que refleja claramente una situación de diseño de cocinas en un espacio determinado.

En cuanto al resto de fenómenos de vocales contiguas que aparecen en los cinco diálogos analizados, se observa que, el caso de las sinalefas, a pesar de ser muy frecuente, resulta un tema muy extenso, pues las realizaciones que se generan pueden ser bastantes. Para ello, resultaría mejor crear un diccionario de pronunciación en el que se incluyeran todos los casos posibles de aparición de diptongos en sinalefas. Consecuentemente, en los modelos del lenguaje podría incluirse la sinalefa como una palabra del diccionario y su posible combinación con otra palabra. Por ejemplo, el diptongo que se forma en la sinalefa *ahí esta bien*, quedaría de la siguiente manera, [ajes*ʃ*tá] + [bjén].

En contraste, los hiatos y diptongación de hiatos son fenómenos que no aparecen frecuentemente en este corpus, pero dada su gran presencia en el habla del español de México es importante su estudio. En la imagen 13 se muestra una gráfica, donde se observa la diferencia que existe entre la frecuencia de cada uno de estos fenómenos.

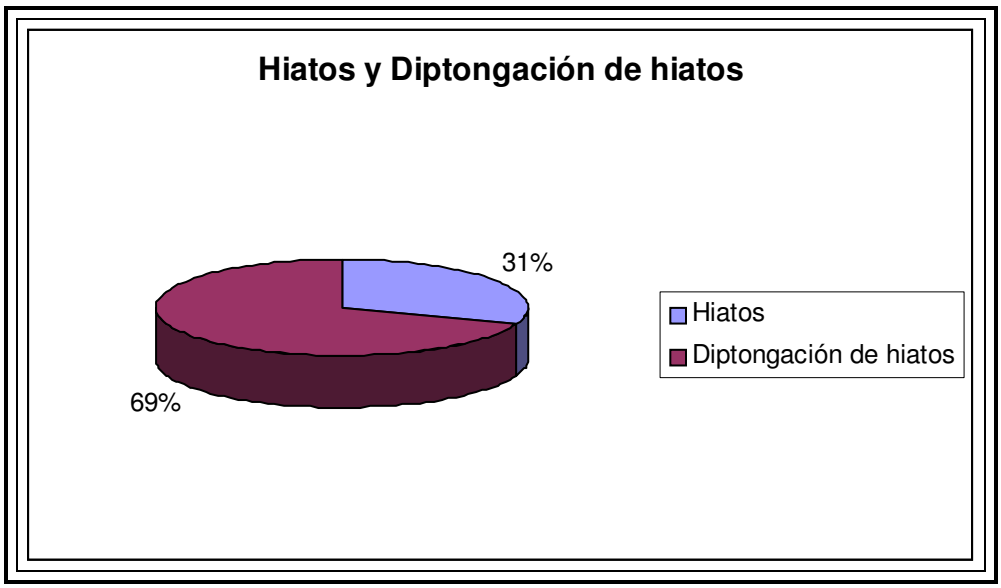

Imagen 13. Frecuencia de aparición de los hiatos y del fenómeno de diptongación de hiatos en cinco de los 31 diálogos del Corpus DIME

Se observa que los hiatos son menos frecuentes que la diptongación de éstos. Uno de los hiatos que constantemente se diptonga es el que se presenta en la palabra *ahí*, refiriéndose al adverbio de lugar, así como también el hiato de la palabra *ahora*, que aunque un poco menos frecuente, tiene un alto grado de diptongación.

Esta diptongación se puede deber al hecho de que el habla que existe en el Corpus DIME es espontánea, rápida y en forma de diálogo.

No obstante, a pesar de que dentro de los hiatos la diptongación es mucho más frecuente, comparando estos datos con el resto de apariciones de vocales contiguas dentro de una palabra en el Corpus DIME (diptongos), encontramos que este fenómeno no es tan común, pues ocupa solamente el 9.59 % del total.

Sin embargo, sería de gran interés realizar un estudio de la diptongación de hiatos en el español de México.

5.4. Resultados

Al igual que el *Esbozo de una nueva gramática de la lengua española* que dice que "todas las vocales pueden combinarse en los diptongos crecientes: *li-diar, li-dié, li-dió, a-guar, a-güé, a-guó*" (Alarcos, 1950:48), y que estos diptongos crecientes son mucho más frecuentes en español que los decrecientes, el corpus analizado en este trabajo encuentra que los diptongos crecientes son los más comunes dentro de la lengua. En cuanto a si los 449 diptongos son crecientes o decrecientes, acentuados o inacentuados y en sílaba abierta o cerrada, en el cuadro 10 se observan las siguientes cifras:

Acentuados, abiertos, decrecientes	**723 diptongos**
Acentuados, abiertos, crecientes	**950 diptongos**
Acentuados, cerrados, decrecientes	**578 diptongos**
Acentuados, cerrados, crecientes	**805 diptongos**
Inacentuados, abiertos, decrecientes	**542 diptongos**
Inacentuados, abiertos, crecientes	**769 diptongos**
Inacentuados, cerrados, decrecientes	**397 diptongos**
Inacentuados, cerrados, crecientes	**624 diptongos**

Cuadro 11. Distribución de los diptongos en 5 de los 31 diálogos del Corpus Dime, según sus características

Comparando estos resultados con los obtenidos por Salporta y Cohen (1958:373) en un corpus basado en 6, 702 palabras in Buchanan's *A graded Spanish word book,* de las cuales obtuvieron un total de 1, 596 diptongos, se observan, en el cuadro 11, los siguientes datos:

Clase de diptongos	Corpus DIME	A graded Spanish word book
Acentuados, abiertos, decrecientes	**723 = 13.41 %**	**45 = 2.81 %**
Acentuados, abiertos, crecientes	**950 = 17.63 %**	**197 = 12.34 %**
Acentuados, cerrados, decrecientes	**578 = 10.72 %**	**15 = 0.93 %**
Acentuados, cerrados, crecientes	**805 = 14.94 %**	**712 = 44.61 %**
Inacentuados, abiertos, decrecientes	**542 = 10.05 %**	**85 = 5.32 %**
Inacentuados, abiertos, crecientes	**769 = 14.27 %**	**509 = 31.89 %**
Inacentuados, cerrados decrecientes	**397 = 7.36 %**	**9 = 0.56 %**
Inacentuados, cerrados, crecientes	**624 = 11.58**	**24 = 1.50 %**

Cuadro 12. Comparación de la distribución de diptongos en dos corpus según sus características

La mayor diferencia se encuentra entre los diptongos acentuados, abiertos, crecientes y los acentuados, cerrados, crecientes, ya que a pesar de que los diptongos -tanto acentuados como crecientes- son los más comunes en ambos corpus, hay una gran diferencia de aparición en cuanto a los cerrados y abiertos; mientras que los abiertos son los más frecuentes en este trabajo, en el de Saporta y Cohen, son los cerrados.

La otra diferencia que se observa es en los diptongos acentuados, abiertos, decrecientes e inacentuados, abiertos, decrecientes, ya que en el Corpus DIME son más frecuentes los primeros que los segundos. En el trabajo de Saporta y Cohen consideran los segundos más frecuentes que los primeros.

Fuera de estas dos diferencias, el resto de los resultados presentan una equivalencia entre las frecuencias de diptongos obtenidos en ambos trabajos. Cabe señalar que hay una diferencia en la obtención de los corpus: mientras que uno es oral, el otro se obtuvo de manera escrita, lo cual seguramente es una de las razones por la cual los datos difieren un poco.

De esta forma, se advierte que, en el caso específico del español, la presencia de diptongos es alta, por lo que la transcripción fonética computacional que se muestra en la imagen 14, y que es la propuesta de esta investigación, sería muy práctica.

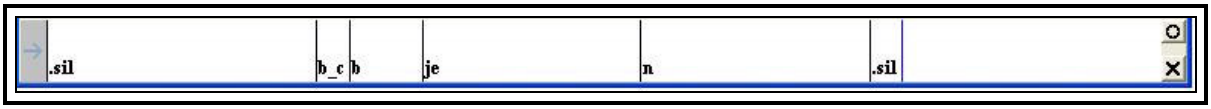

Imagen 14. Nueva transcripción fonética computacional de los diptongos

En cuanto a las sinalefas, se observa que, al igual que los diptongos, aparecen en gran cantidad, por lo que se deben tomar en cuenta para la nueva transcripción fonética computacional. En la imagen 15 se observa una etiqueta de sinalefa formando diptongo transcrita con la nueva etiquetación.

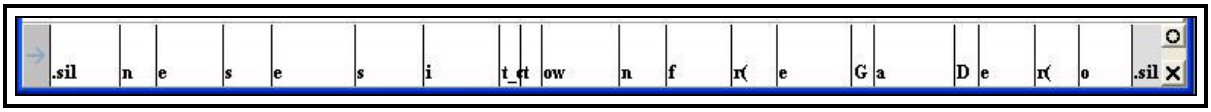

Imagen 15. Nueva transcripción fonética computacional de las sinalefas

Con respecto al diccionario de pronunciación, en el cuadro 16 se muestra cómo quedaría la transcripción fonética computacional para los diptongos y para las sinalefas.

Diptongos	Sinalefas
b_c b je n t_c t je n e s m we V e s t_c t r(aj m e	a s je s t_c t a n e s e s j t_c t ow n a o r(aw n m a k_c k i n aj f r(e G a D e r(o

Cuadro 13. Nueva transcripción fonética computacional para el diccionario de pronunciación

Con respecto a los modelos del lenguaje, en el cuadro 13 se ejemplifica cómo quedaría la transcripción fonética computacional de las sinalefas. El caso de los diptongos no tiene cambio, excepto por los nuevos modelos acústicos que lo forman.

Palabra	Posibles combinaciones
n e s e s j t_c t ow n	**f r(e G a D e r(o** **m we V l e** **l a V a t_c t r(a s t_c t e s**

Cuadro 14. Nueva transcripción fonética computacional de las sinalefas formando diptongos, para los modelos del lenguaje

Finalmente, los hiatos y la diptongación de hiatos son fenómenos que, aunque no son el tema de esta tesis, es importante mencionarlos, porque al igual que los diptongos son dos vocales contiguas dentro de una palabra, sea o no en la misma sílaba.

Como opinan dos autores consultados, los hiatos suelen ser muy frecuentes en la lengua hablada en México por su recurrente diptongación. Perissinotto dice: "debo señalar que la diptongación es la regla más que la excepción en este dialecto: en 236 casos de hiatos etimológicos, aparecieron 170 diptongos y sólo se mantuvieron los hiatos en 66 ocasiones" (Perissinotto, 1975:39) y a partir del estudio exhaustivo, hecho por Alonso, para explicar el fenómeno de la tendencia antihiática que existe en el habla, Matluck concluye que "el fenómeno

1. **es característico del habla vulgar de España y de América, con pocas excepciones;**
2. **no es importado, sino que se produjo en América misma, como evolución popular de la lengua;**
3. **ha penetrado, tanto en España como en América, en todas las clases sociales:**
4. **es más común hoy día en el lenguaje culto de América que en el de España, pero esto se debe solamente a que la reacción culta llegó más tarde a América que a España" (1951:38).**

Los resultados que se obtienen en el Corpus DIME también reflejan esta tendencia, pues a pesar de que la aparición de hiatos es muy baja frente a la de diptongos, la gran mayoría de éstos se diptonga.

En este capítulo se dio una pequeña reseña de lo que constituye el Proyecto DIME, específicamente del corpus que se utiliza para crear los modelos del lenguaje.

Posteriormente, se analizó la frecuencia de aparición de los fenómenos conocidos como diptongos, sinalefas, hiatos y diptongación de hiatos. Finalmente, se concluyó que la aparición de diptongos es la mas frecuente, siguiendo el caso de las sinalefas, para concluir con la diptongación de hiatos y la de hiatos. Estos dos últimos poco frecuentes, sin embargo el primero con más realizaciones que el segundo.

6. CONCLUSIONES

El propósito de esta tesis de licenciatura fue proponer una nueva transcripción fonética para un corpus que tenga la finalidad de crear los modelos acústicos para un reconocedor de habla. Esta propuesta tiene como objetivo transcribir los diptongos como una unidad y no como dos, como hasta ahora se ha hecho. Para entender el impacto que esto tendría en la arquitectura de un reconocedor de habla, en una primera parte se hizo un breve recorrido por la historia de la computación. Teniendo esto como antecedente, se habló en concreto de las tecnologías del habla, haciendo la diferencia entre síntesis de habla y reconocimiento de habla, para finalmente describir la arquitectura de un reconocedor de habla; es decir, el proceso que se lleva a cabo en el reconocimiento del lenguaje.

De esta manera, se estableció el lugar donde la transcripción monofonemática de los diptongos llevaría a un cambio en la arquitectura de un reconocedor de habla. Así, a los modelos acústicos que se tengan, deberían aumentarse 14 modelos más, correspondientes a los 14 diptongos que se presentan en el español de México. Consecuentemente el diccionario de pronunciación debería contener esta nueva información. Cabe mencionar que la posible diptongación de vocales fuertes representa un estudio aparte del que se hizo en este trabajo, por lo que no se incluyen los modelos acústicos de los mismos.

Con el fin de entender de qué manera el lingüista aporta sus conocimientos en el proceso de construcción de un reconocedor de habla, en el siguiente capítulo se proporcionó una breve descripción de cómo se lleva a cabo la transcripción fonética de un

corpus. Para ello, fue necesario introducir el término de fonética instrumental, tomando como ejemplo específico el uso del espectógrafo. Además, se mencionaron los alfabetos fonéticos más utilizados hoy en día y los alfabetos computacionales que de ellos se desprenden, así como también se enumeraron algunas de las técnicas que se utilizan para crear un corpus oral. Se mencionó el caso particular de los experimentos "Mago de Oz" por ser la técnica en la que se basó la elaboración del corpus que se trabajó en esta tesis, el Corpus DIME.

Era de gran importancia para el trabajo explicar en qué consiste la etapa de la transcripción fonética pues, además de ser una labor lingüística, es la primera fase en la que la propuesta de considerar a los diptongos como monofonemáticos produce ya un cambio.

Para saber lo que dice la teoría de los diptongos, en el capítulo posterior se hizo una clasificación de las vocales, con la finalidad de introducir los términos diptongo e hiato. Se entendió por diptongo la presencia de dos vocales en una misma sílaba, y además se mencionó, en el caso de las sinalefas, dos vocales que se unen a causa del enlace entre dos palabras y que forman una sola sílaba. El hiato es, por el contrario, la presencia de dos vocales contiguas pero en diferente sílaba. Sin embargo, el hiato suele diptongarse con mucha frecuencia, por lo que también fue preciso mencionar el fenómeno de diptongación de hiatos. Por último, se contrastaron las ideas de dos autores, Trubetzkoy y Alarcos, de considerar a los diptongos como monofonemáticos en el caso del primero, y como bifonemáticos en el caso del segundo.

Finalmente, fue necesario observar si los fenómenos que de dos vocales contiguas se desprenden, es decir los diptongos, las sinalefas, los hiatos y la diptongación de hiatos, son tan frecuentes en la lengua como para considerar relevante la propuesta de esta tesis.

Para esto, en el último capitulo se explicó en qué consiste el Proyecto DIME, del que forma parte este trabajo, así como el Corpus DIME que se utilizó en el análisis de los datos.

Los resultados que se obtuvieron mostraron que tanto los diptongos como las sinalefas tienen una gran presencia en el Corpus DIME, siendo los hiatos y la diptongación de hiatos mucho menos frecuentes. Sin embargo, comparando únicamente los datos obtenidos de estos dos últimos fenómenos resulta mucho más frecuente el segundo que el primero.

Así, a partir de los resultados obtenidos y de los datos que éstos arrojan, al ser los diptongos un fenómeno frecuente en el corpus, su transcripción monofonemática podría ayudar al reconocimiento de habla. Si esto se llevara a cabo, los modelos acústicos del reconocedor aumentarían y la información del diccionario de pronunciación también tendría que ser ampliada a partir de los nuevos modelos acústicos. Esto para el caso de los diptongos; en el caso de las sinalefas que forman diptongo, lo que cambiaría en la arquitectura del reconocedor, además de los modelos acústicos, sería, al igual que en los diptongos, el diccionario de pronunciación que debería incluir todas las posibles combinaciones de palabras del corpus donde sería factible la formación de un diptongo. Además, los modelos del lenguaje tendrían que considerar a la sinalefa como una sola palabra y, por lo tanto, buscar todas las posibles combinaciones de palabras que siguieran a ésta.

De esta manera, esta tesis abundó en los temas de interés conjunto para la lingüística y para la computación, unidas a partir de la lingüística computacional. Se abren aquí, pues, bases teóricas para trabajos futuros que tenga el interés de realizar un trabajo práctico para la resolución de los segmentos vocálicos contiguos en los sistemas de reconocimiento de habla para el español de México

7. REFERENCIAS BIBLIOGRÁFICAS

ACERO, ALEJANDRO. 1995 (Ms.). "The role of phoneticians in speech technology", en *European studies in phonetics and speech communication*, G. Bloothooft, V. Hazan, D. Huber y J. Llisterri (Eds.), Essex: OTS Publications.

AGUILAR, LOURDES. 2003 "Effects of segmental and prosodic variables on vowel sequences pronunciation in Spanish", en *Proceedings of the 15th International Congress of Phonetic Sciences.* Barcelona, 3-9 August 2003, M. J. Solé, D. Recasens and J. Romero (Eds.), pp. 2111- 2114.

AGUILAR, LOURDES, JOAQUIM LLISTERRI BOIX y MARÍA JESÚS MACHUCA. 1994 (Ms.). "Some phonetic data on speech produced in different experimental situations", ponencia presentada en *ESPRIT BRA VOX Workshop*, Orsay: LIMSI.

ALARCOS LLORACH, EMILIO. 1950/1991. *Fonología española*, Madrid: Gredos.

BLAHETA, DON. Ms. "Representation of IPA with ASCII", Providence: Brown University, Department of Computer Science. [Cito por: http://www.blahedo.org/ascii-ipa.html]

CUÉTARA PRIEDE, JAVIER. 2004 (Ms.). *Fonética y fonología del habla espontánea de la ciudad de México. Su aplicación en las tecnologías del habla*, tesis de maestría inédita, México: Universidad Nacional Autónoma de México.

DAHLBÄCK, NILS, ARNE JÖNSSON y LARS AHRENBERG. 1993. "Wizard of Oz studies – Why and how", en *Intelligent users interfaces*, Linköping: Laboratorio de Procesamiento de Lenguaje Natural, pp. 193-200.

GILI GAYA, SAMUEL. 1975. *Elementos de fonética general*, Madrid: Gredos.

HIERONYMUS, JAMES L. 1994 (Ms.). "ASCII phonetic symbols for the world's languages: Worldbet", Nueva Jersey.

————. 1997 (Ms.). "Worldbet phonetic symbols for multilanguage speech recognition and synthesis", Nueva Jersey.

INTERNATIONAL PHONETIC ASSOCIATION. 1949/1971. *The principles of the International Phonetic Association*, Londres: University College.

————. 1999. *Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet*, Cambridge: Cambridge University Press.

JURAFSK, DANIEL y JAMES H. MARTIN. 2000. *Speech and Language Processing*, New Jersey: Prentice-Hall.

LANDER, TERRI. 1997. *The CSLU labeling guide*, Oregon: Oregon Graduate Institute of Science and Technology.

LLISTERRI BOIX, JOAQUIM. 1991. *Introducción a la fonética: el método experimental*, Barcelona: Anthropos.

————. 1997 (Ms.). "Transcripción, etiquetado y codificación de corpus orales", en *Etiquetación y extracción de información de grandes corpus textuales*, curso impartido en *Seminario de Industrias de la Lengua*, Soria.

————. 2003. “Las tecnologías del habla”, en *Tecnologías del lenguaje,* María Antonia Martí (Coord.), Barcelona: UOC, pp. 249-281.

LLISTERRI BOIX, JOAQUIM, LOURDES AGUILAR, JUAN M. GARRIDO, MARÍA JESÚS MACHUCA, RAFAEL MARÍN, CARME DE LA MOTA Y ANTONIO RÍOS. 1999. “Fonética y tecnologías del habla”, en *Filología e Informática. Nuevas tecnologías en los estudios filológicos,* J.M. Blecua, G. Clavería, C. Sánchez y J. Torruella (Eds.), Barcelona: Milenio i Universitat Autónoma de Barcelona, pp. 449-479.

LLISTERRI BOIX, JOAQUIM, CARMEN CARBÓ, MARÍA JESÚS MACHUCA, CARME DE LA MOTA, MONTSERRAT RIERA y ANTONIO RÍOS. 2003 (Ms.). “El papel de la lingüística en el desarrollo de las tecnologías del habla”, en *Séptimas Jornadas de Lingüística*, Cádiz: Servicio de Publicaciones de la Universidad de Cádiz.

LLISTERRI BOIX, JOAQUIM, JUAN M. GARRIDO. 1998. “La ingeniería lingüística en España”, en *El español en el mundo*. Anuarios del Instituto Cervantes.

LLISTERRI BOIX, JOAQUIM y JOSÉ B. MARIÑO. 1993 (Ms.). “Spanish adaptation of SAMPA and automatic phonetic transcription”, Barcelona.

LLISTERRI BOIX, JOAQUIM, MARÍA ANTONIA MARTÍ. 2002. “Las tecnologías lingüísticas en la sociedad de la información”, en *Tratamiento del lenguaje natural,* M. A. Martí y J. Llisterri (Eds.), Barcelona: Fundación Duques de Soria y Edicions Universitat de Barcelona, pp. 13-28.

LOPE BLANCH, JUAN M. 1969/1989. *La filología hispánica en México. Tareas más urgentes*, México: Universidad Nacional Autónoma de México.

MADRID SERVÍN, EDGAR A. y MARIO A. MARÍN RODRÍGUEZ. 2001. “Estructuras formánticas de las vocales del español de la ciudad de México”, en *Temas de fonética instrumental*, E. Herrera (Ed.), México: El Colegio de México, pp. 39-58

MARTÍNEZ CELDRÁN, EUGENIO. 1984. *Fonética (con especial referencia a la lengua castellana)*, Barcelona: Teide.

MARTÍNEZ CELDRÁN, EUGENIO, ANA MA. FERNÁNDEZ PLANAS y JOSEFINA CARRERA-SABATÉ. 2003. “Illustrations of the IPA. Castilian Spanish”, *Journal of the International Phonetic Association*, 33:2, pp. 255-260.

MATLUCK, JOSEPH. 1951. *La pronunciación en el español del Valle de México*, México: Edición de autor.

MORENO DE ALBA, JOSÉ G. 1994. “Las vocales”, en *La pronunciación del español en México*, México: El Colegio de México. pp. 31-63.

MOTA, CARME DE LA y ANTONIO RÍOS. 1995. “Problemas en torno a la transcripción fonética del español: los alfabetos fonéticos propuestos por IPA y *RFE* y su aplicación a un sistema automático”, en *Acta Universitatis Wratislaviensis*, Wroclaw: Universidad de Wroclaw, pp. 97-109.

NAVARRO TOMÁS, TOMÁS. 1918/1982. *Manual de pronunciación española*, Madrid: RAYCAR.

————. 1944/1966. *Manual de entonación española*, México: Colección Málaga.

OLIVIER, MARÍA ALEJANDRA. 1999. *Evaluación de métodos de determinación automática de una transcripción fonética*, tesis de maestría inédita, México: Universidad de la Américas.

PÉREZ, HERNÁN EMILIO. 2003 (Ms.). “Frecuencia de fonemas”, Concepción: Universidad de Concepción.

PERISSINOTTO, GIORGIO SABINO ANTONIO. 1975. *Fonología del español hablado en la Ciudad de México. Ensayo de un método sociolingüístico*, México: El Colegio de México.

PINEDA, LUIS A., ANTONIO MASSÉ, IVÁN MEZA, MIGUEL SALAS, ERIK SCHWARZ, ESMERALDA URAGA y LUIS VILLASEÑOR. 2001 (Ms.). "El Proyecto DIME", México: Universidad Nacional Autónoma de México.

PINEDA, LUIS A., LUIS VILLASEÑOR, JAVIER CUÉTARA PRIEDE, HAYDE CASTELLANOS e IVONNE LÓPEZ. 2004 (Ms.). "DIMEx100: A new phonetic and speech corpus for Mexican Spanish", México: Universidad Nacional Autónoma de México.

QUILIS ANTONIO. 1981/1988. *Fonética acústica de la lengua española*, Madrid: Gredos.

————. 1985. *El comentario fonológico y fonético de textos*, Madrid: Arco Libros.

————. 1993/1999. *Tratado de fonología y fonética españolas*, Madrid: Gredos.

REAL ACADEMIA ESPAÑOLA. 1973/1995. *Esbozo de una nueva gramática de la lengua española*, Madrid: Espasa Calpe.

REVISTA DE FILOLOGÍA ESPAÑOLA. 1915. "Alfabeto fonético", *Revista de Filología Española*, Madrid, 2, pp. 374-376.

SALPORTA, SOL. 1956. "A note on Spanish semivowels", en *Language*, 32, pp. 287-290.

SALPORTA, SOL y RITA COHEN. 1958. "The distribution and relative frequency of Spanish diphtongs", en *Romance Philology*, 40, pp. 371-377.

SOLÉ SABATER, MARIA-JOSEP. c. 1985. "La experimentación en fonética y fonología", en *Estudios de fonética experimental*, E. Martínez Celdrán y M. J. Solé Sabater (Eds.), Barcelona: Promociones Publicaciones Universitarias, pp. 1-70

TAPIAS MERINO, DANIEL. 2002. "Interfaces de voz con lenguaje natural", en *Tratamiento del lenguaje natural,* María Antonia Martí, Joaquim Llisterri Boix (Eds.), Barcelona: Edicions Universitat de Barcelona, pp. 189-208.

TLATOA. 2000. "How to hand label a speech corpus", México: Universidad de las Américas. [Cito por:http://mailweb.udlap.mx/~sistemas/tlatoa/howto/hand_label.html]

TRUBETZKOY NIKOLAI. SERGEEVICH. 1964/1973. *Principios de fonología*, Madrid: Cincel.

URAGA, ESMERALDA y LUIS A. PINEDA. 2000 (Ms.). "A set of phonological rules for Mexican Spanish", México: Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas.

VAQUERO DE RAMÍREZ, MARÍA. 1996. *El español de América I: pronunciación,* Madrid: Arco Libros.

VILLASEÑOR, LUIS, ANTONIO MASSÉ y LUIS A. PINEDA. 2001 (Ms.). "The DIME Corpus", en *Tercer Encuentro Internacional de Ciencias de la Computación ENC-01*, Aguascalientes: SMCC-INEGI, pp. 591-600.