Diseño inicial de un instrumento fonético-acústico para evaluar la familiaridad de los sujetos con unidades léxicas[*]

Initial design of a phonetic-acoustic instrument to test subjects' lexical unit knowledge

María Natalia Castillo Fadić
Pontificia Universidad Católica de Chile
mcastilf@uc.cl

Pilar Oplustil
University of Edinburgh
s1520337@sms.ed.ac.uk

Resumen

El presente proyecto se propone la construcción de un instrumento que, por medio de la aplicación de herramientas de lingüística computacional y fonética acústica, evalúe si los hablantes conocen o no unidades léxicas dadas. El instrumento, en etapa de desarrollo inicial, podría ser aplicado a distintos conjuntos léxicos, en investigaciones diversas. Como producto esperado, se obtendrá un programa computacional afinado especialmente para el estudio del español de Chile.

Palabras clave: lexicología, reconocimiento léxico, fonética acústica, lingüística computacional.

Abstract

In this project, we propose to build an instrument to evaluate if a speaker knows certain lexical units, applying tools from computational linguistics and acoustic phonetics. The instrument, which is a prototype, could be applied to different lexical clusters, in various types of research. We expect as a result to end up with an application specifically tuned to study Chilean Spanish.

Keywords: lexicology, lexical recognition, acoustic phonetics, computational linguistics.

[*] Proyecto de Investigación INTERDISCIPLINA Nº 1/2014 VRI UC (2014-2016)

1. INTRODUCCIÓN

Este estudio se propone profundizar en la evaluación automática de la familiaridad de los hablantes con unidades léxicas específicas, a partir de la aplicación de herramientas de lingüística computacional y fonética acústica al procesamiento y análisis de la lectura en voz alta de listas de palabras.

En esta línea, encontramos antecedentes en Wright (1979) y en Davies *et al*. (2013). El primero realizó un experimento en que midió la tasa de lectura en voz alta de una lista de palabras en inglés de una sílaba, donde clasificó algunas como raras (frecuencia absoluta o $f_i \leq 3$ sobre un millón) y otras como comunes ($f_i \geq 100$ sobre un millón). Generó tres sets para cada tipo de palabras, de tres, cuatro y cinco letras (Wright, 1979: 412); observó que las clasificadas como raras mostraron mayor duración que las consideradas comunes, aun cuando tenían el mismo número de letras: las diferencias oscilaron entre 18 y 54 ms., es decir, entre un 17% y un 24% de la duración total de la palabra (Wright, 1979: 417). El segundo estudió los factores que intervienen en la fluidez de la lectura en voz alta en niños hispanohablantes de distintas edades para determinar las diferencias entre los que presentaban dislexia y los que no; midió tanto el tiempo de respuesta (la duración de emisión de la palabra) como el tiempo de reacción (Davies *et al*., 2013: 725) ante una lista conformada por palabras escogidas según su f_i de aparición en textos de lectura obligatoria y pseudopalabras (Davies *et al*., 2013: 728): observó que el tiempo de reacción fue más corto para palabras más cortas y más frecuentes; los resultados fueron similares para el tiempo de la respuesta: las palabras más cortas y más frecuentes mostraron menor duración (Davies *et al*., 2013: 734); sin embargo, a pesar de lo que señala la literatura, la influencia del largo intrínseco de la palabra no varió según la edad de los niños; y, por otro lado, la influencia de la f_i de la palabra fue menor para el grupo de niños mayores. Según el autor, esto podría deberse a: "the gradual optimization of connection weights, a process of adaptation to experience that tends to narrow the space in which the frequency effect can appear" (Davies *et al*., 2013: 735).

Hasta el momento no hemos encontrado estudios que analicen la relación entre familiaridad léxica y entonación de la emisión.

Generar un instrumento de medición centrado en el nivel léxico es el objetivo del proyecto que aquí presentamos. Este instrumento, en su concepción general, podrá ser aplicado a la evaluación automática de la familiaridad de los hablantes con cualquier conjunto léxico, por cuanto las metodologías que aplica y los supuestos teóricos que lo sostienen son independientes del área temática en estudio. No obstante, las listas de palabras sobre las cuales se opere, dependerán de las necesidades específicas de cada investigación y de los conjuntos léxicos de interés para cada una.

El éxito del instrumento dependerá de la previa comprobación para el español de Chile de una serie de hipótesis que suponen la existencia de una relación significativa entre familiaridad léxica y variables suprasegmentales. En cuanto a la relación entre familiaridad de los sujetos con unidades léxicas y los valores suprasegmentales que evaluaremos a través del instrumento, esperamos encontrar los siguientes comportamientos: (1) A mayor

familiaridad, la lectura en voz alta será más fluida, es decir, más veloz o de menor duración; (2) A menor familiaridad, la lectura en voz alta será menos fluida, es decir, menos veloz o de mayor duración; (3) A mayor familiaridad, la lectura en voz alta presentará una entonación enunciativa o enumerativa; (4) A menor familiaridad, la lectura en voz alta podría presentar una entonación interrogativa; (5) A mayor familiaridad, el tiempo de reacción entre un término y otro será menor; (6) A menor familiaridad, el tiempo de reacción entre un término y otro será mayor.

A continuación, presentamos nuestros fundamentos teóricos y metodológicos y damos cuenta del proceso de desarrollo del instrumento y de su grado de avance.

2. LEXICOLOGÍA Y FONÉTICA: LA PROSODIA COMO REFLEJO DE LA FAMILIARIDAD LÉXICA

2.1. Nivel léxico

El nivel léxico constituye el núcleo a partir del cual se expresan las significaciones de una lengua. Su carácter de significador e interpretador del mundo explica su gran permeabilidad a la cosmovisión de los hablantes, lo que lo hace el nivel más sensible a variaciones diacrónicas, diatópicas, diastráticas y diafásicas, así como a diferencias de registro.

Si el *Corpus Básico del Español de Chile* (Castillo Fadić, 2012a) contiene más de cincuenta y tres mil vocablos diferentes sobre un total de más de quinientas mil palabras, repartidas en cinco mundos (Drama, Narrativa, Ensayo, Técnico-Científico y Prensa), menos de cinco mil forman parte de nuestro *léxico básico* (cf. Castillo Fadić, 2012b, 2015), es decir, del léxico atemático de mayor uso en la comunidad. Este léxico básico, pese a su relevancia, por el hecho de obtenerse a partir de las unidades léxicas de mayor uso (U), donde el U se define como el producto de la frecuencia (f_i) por la dispersión (D), necesariamente deja fuera a las unidades léxicas fuertemente temáticas, así como a las circunscritas a situaciones o a disciplinas específicas. Para la medición del léxico temático, no pueden usarse meros cálculos estadísticos que observen la f_i de realización de unidades léxicas en enunciados orales o escritos, sino que es preciso realizar mediciones indirectas, mediadas. Una de estas mediciones es la de la *disponibilidad léxica*, basada en la activación del léxico que los sujetos conservan en su memoria pasiva, y que sólo emplean cuando la situación comunicativa lo amerita. El léxico disponible, o léxico de mayor disponibilidad, está compuesto por las unidades temáticas del lexicón mental de los miembros de una comunidad que se activan con mayor facilidad cuando la situación comunicativa se desarrolla en torno a un tema determinado; por lo mismo, está compuesto preferentemente por formas nominales, seguidas por verbos de alta especificidad semántica. Tanto el léxico básico como el disponible se obtienen a partir de la aplicación de fórmulas estadísticas a corpus de referencia etiquetados: en el primer caso, a corpus obtenidos de textos reales, pertenecientes a distintos mundos o géneros discursivos; en el segundo, a corpus obtenidos de la aplicación de encuestas escritas a informantes, a los que se les dan dos minutos para escribir palabras en relación con temas o centros de interés propuestos por un investigador. Juntos conforman el léxico fundamental de una comunidad, aquel que se necesita para

desenvolverse de manera adecuada en distintas instancias comunicativas. En distintas lenguas, las unidades léxicas ubicadas en los primeros rangos de U son palabras gramaticales, seguidas por unidades de baja especificidad semántica, estructura fónica simple y extensión breve. En el caso del español, esto es lo que observa Ávila (1998: 256) en Málaga–"siguiendo con la tónica observada en la mayoría de las lenguas, las palabras de más uso en nuestro corpus son cortas y poco complejas fonéticamente, de tal manera que más de la cuarta parte son monosílabos y el resto bisílabos a excepción de algunos elementos con tres sílabas"– y lo que ratifica Castillo Fadić (2012b) en relación con el español de Chile.

Los vocablos de mayor disponibilidad no presentan necesariamente una alta f_i, ni menos aún, un alto U, pero pueden resultar tan familiares como las unidades léxicas de alta f_i o alto U. Esto explica que vocablos como <velador>, "mesa de noche," o <elefante>, "cierto mamífero paquidermo", puedan aparecer infrecuentemente en textos escritos y orales (a menos de que estos traten específicamente sobre muebles del dormitorio o animales salvajes), y aun así corresponder a vocablos familiares para los hablantes, en virtud de su disponibilidad.

La natural tendencia de la lengua a la simplificación redunda en que, existiendo dos variantes para una misma variable léxica, tenga mayores posibilidades de prevalecer aquella cuya realización sea más sencilla. Esto explica que unidades léxicas con un número de fonemas sobre la media, composición fonémica infrecuente y presencia de fonemas complejos, tiendan a ser menos frecuentes que formas equivalentes o cuasi equivalentes de menor complejidad. Por otra parte, la familiaridad de los hablantes con una unidad léxica no depende sólo del conocimiento que tengan de esa misma unidad, sino también de la posibilidad de incluirla en una familia de palabras en sentido amplio, regida por afinidad fonosemántica. Esto significa que las unidades léxicas cuyo significado y estructura sean abiertamente disímiles a todas aquellas que el hablante conoce, tenderán a resultarle menos familiares y sencillas de recordar e incorporar que aquellas que, siendo también desconocidas, presenten una estructura fonosemántica similar a la de otras unidades léxicas conocidas, a cuya familia léxica el hablante pueda adscribirlas. Esto explica incluso que algunas unidades puedan caer en desuso, como expone Castillo Fadić (2001).

2.2. Nivel segmental y nivel suprasegmental

El sistema fonético y fonológico de toda lengua se compone de un nivel *segmental* y de un nivel *suprasegmental*. El segmental corresponde al conjunto de fonemas vocálicos y consonánticos con todos los rasgos que los caracterizan. El suprasegmental, por su parte, está por sobre estas unidades, y consiste en propiedades que abarcan uno o más segmentos en toda la cadena fónica. Estas propiedades son, entre otras, la duración, la velocidad y la frecuencia fundamental.

2.2.1 La *duración* corresponde al tiempo medible en segundos o milisegundos desde que una unidad comienza hasta que termina. El control de la duración de una emisión articulada se produce a través de la actividad de la cavidad infraglótica. La duración de una palabra depende del número de segmentos que la conforman; a su vez, cada uno de estos segmentos posee una duración propia, determinada tanto por su modo de articulación, como por el

punto de articulación (Gil, 2007: 65) y por la coarticulación entre segmentos adyacentes (Fernández Planas, 2011: 55). En promedio, los segmentos pueden durar entre 30 y 300 ms. en una lengua como el español (Gil, 2007: 64). En un nivel suprasegmental, el estudio de la duración se ha centrado en la sílaba más que en la palabra o el grupo fónico. Las sílabas tónicas, por ejemplo, poseen mayor duración que las átonas, duración que aumenta si la sílaba tónica se encuentra en el tonema de la frase (Fernández Planas, 2011: 56). Sin embargo, el nivel de familiaridad que el hablante tiene con una palabra puede también afectar la duración con que esta es emitida: "the duration of words in prose passages read aloud (controlled for the number of syllables and phrase boundary type, but not for number of letters) decreases with increasing frequency" (Wright, 1979: 412). Es decir, mientras más familiar sea una palabra, la duración de su emisión es menor, y viceversa.

2.2.2. La duración con que se emite una palabra o una cadena fónica tiene relación directa con la *velocidad del habla* y, en el caso de la lectura, con la *fluidez de lectura en voz alta*. La velocidad del habla, también conocida como velocidad de elocución o tempo, se relaciona con "la mayor o menor rapidez con que un hablante pronuncia sus enunciados" (Gil, 2007: 308). Se suele medir contando un cierto número de unidades producidas en un período de tiempo determinado: palabras, sílabas o segmentos. Mediciones de la velocidad del habla en español peninsular señalan que la velocidad normal se caracteriza por 205 palabras por minuto (Wainschenker *et al*., 2002: 2); en inglés, en tanto, la velocidad normal de habla estaría en un promedio de 200 palabras por minuto (Wainschenker *et al*., 2002: 2). La velocidad de la lectura en voz alta también se ha tendido a analizar de esta manera. Wainschenker *et al*. (2002) analizan la velocidad de habla midiendo alófonos por segundo en el español rioplatense a través de la lectura de textos a diferentes velocidades: según sus resultados, la lectura de un texto a velocidad normal tuvo como promedio 12,9 alófonos por segundo; la lectura lenta promedió 3,3 a/s; y la lectura rápida, 17,7 a/s. (2002: 3). Por otra parte, la fluidez de lectura oral se puede definir como "the ability to read text quickly, accurately, with proper phrasing and expression, thereby reflecting the ability to simultaneously decode and comprehend" (Valencia *et al*., 2010: 271).

Fluidez y velocidad de lectura apuntan a conceptos similares, pero difieren en el aspecto correctivo: la fluidez no se estudia solamente midiendo el número de unidades en un período (como la velocidad del habla), sino notando si estas unidades han sido correctamente pronunciadas o no, es decir, lo que se busca es medir "palabras correctas por minuto" (Valencia y otros, 2010: 271). Claramente, el uso más común de esta medición responde a evaluar las habilidades de escolares, tratando de diagnosticar problemas de comprensión lectora; se espera que la medición de palabras correctas por minuto refleje la comprensión de las palabras así como de niveles superiores textuales (Valencia *et al*., 2010: 271), ya que la comprensión depende de que el tiempo de decodificación o acceso léxico sea lo suficientemente rápido para no entorpecer la comprensión (Compton & Carlisle, 1994: 2).

Sin embargo, contar el número de unidades pronunciadas correctamente por minuto no sería una medición pertinente de fluidez de lectura oral para todas las lenguas. La situación puede diferir entre una lengua y otra según la relación que existe entre el sistema fonológico y el sistema ortográfico de cada una. Así, por ejemplo, en el caso de sujetos con dislexia: "In an opaque orthography like English, phonological coding errors are a

prominent feature of dyslexia. In a transparent orthography like Spanish, reading difficulties are characterized by slower reading speed rather than reduced accuracy" (Davies *et al*., 2013: 722), debido a que la relación entre fonema-grafema del español es más unívoca que en el caso del inglés (Davies *et al*., 2013: 722). Si extrapolamos esto a hablantes sin dificultades de lectura, los sujetos enfrentados a una unidad léxica desconocida podrían no sufrir contratiempos al momento de pronunciar cada uno de sus segmentos, pero sí podría alterarse la velocidad con que leen dicha unidad.

La fluidez de la lectura en voz alta se mide a través de diferentes instrumentos: lectura de textos, oraciones, listas de palabras, etc. Las listas de palabras tienen la ventaja de que pueden ser objeto de dos tipos de mediciones relacionadas con la velocidad de producción del sujeto: tanto la velocidad de lectura como la velocidad de reacción o tiempo de reacción entre cada elemento de la lista. (Davies *et al*., 2013: 723). La velocidad de reacción o *latencia* es un indicador de automaticidad en la lectura o decodificación de un elemento. Mientras más familiares son los estímulos, el tiempo de reacción decrece (McCormick & Samuels, 1979: 3).

Velocidad y duración son, en conclusión, elementos relacionados. Mientras aumenta la velocidad, decrece la duración. La velocidad con que se lee una palabra puede deberse tanto a la duración propia de esta como a la f_i de aquella unidad léxica. Estudios han demostrado que a medida que los niños crecen la duración propia de la palabra deja de ser el factor más determinante, mientras que la influencia de su f_i, no varía con la edad (Dives, 2013: 724).

2.2.3 El último componente del nivel suprasegmental es la entonación, que implica diferentes componentes acústicos, siendo el más relevante para el español la frecuencia fundamental (F_0), que "manifiesta la melodía de la entonación" (Martínez & Fernández, 2007: 193) con la que interactúan tanto la intensidad como la duración. A nivel fonológico, la curva melódica es interpretada como la entonación de la emisión. Se suele estudiar con respecto al grupo fónico, el enunciado y el acto de habla y, en menor medida, en el nivel léxico. Sin embargo, en la emisión de una lista de palabras se puede esperar que la palabra se comporte con una entonación similar a la que tendría un enunciado que consistiera en una oración gramatical compleja.

La entonación enunciativa se caracteriza por tener un final descendente de la curva melódica, que suele darse en enunciados declarativos completos y al terminar una enumeración (Gil, 2007: 390). Cuando se realiza una aseveración, la entonación desciende al final del enunciado, comportamiento que se ha visto también en el Español de Santiago de Chile (Uribe et al, 2000: 103), aunque cabe destacar que no es consistente en todas las regiones del país (Muñoz-Builes, 2017: 8-9). En la lectura de una lista de palabras se podría dar que los informantes adoptaran una entonación de enumeración que básicamente abarcaría toda la prueba, tomando cada unidad léxica como un elemento enumerado. La entonación de las enumeraciones se caracteriza por ser ascendente al final de cada elemento enumerado (Gil, 2007: 392), exceptuando el último elemento de la lista el cual tiene una forma plana y baja (Ortiz et al, 2010: 256). Por lo tanto, ante la familiaridad con una unidad léxica, la entonación podría ser tanto enunciativa como enumerativa. Los enunciados interrogativos indagativos, es decir, que esperan una respuesta, a diferencia de preguntas irónicas, interrogaciones pronominales o interrogaciones imperativas (Gil, 2007: 391) se

caracterizan por final ascendente (Gil, 2007: 393). En el Español de Chile, las preguntas indagativas de sí o no también finalizan de manera ascendente, mostrando interés de parte del hablante (Ortiz et al, 2010: 262), el cual podría presentarse en caso de poca familiaridad ante la unidad léxica. Sin embargo, un caso especial son las preguntas de sí o no que buscan una confirmación, las cuales presentan curvas melódicas similares a una aseveración, con una curva melódica similar a un trapecio (ascedente, plana, descendente) (Ortiz et al, 2010: 270). Este tipo de entonación podría presentarse en caso de que el lector busque confirmación o aprobación por parte del evaluador al leer la lista de palabras.

3. DISEÑO DEL INSTRUMENTO

3.1. Elaboración de un listado de unidades léxicas meta

3.1.1 Puesto que el instrumento exige la selección previa de un conjunto de unidades léxicas para someterlas a evaluación, lo primero es determinar los modos de realizar esa selección. Si lo que se desea es evaluar si el hablante reconoce o no unidades léxicas de alto U, f_i, D o disponibilidad en una comunidad, la selección de las unidades léxicas por evaluar podría efectuarse a partir de diccionarios no definitorios disponibles para consulta, que brinden la requerida información estadística. Si la finalidad de la investigación, en cambio, es evaluar la familiaridad del sujeto con léxico específico cuyo conocimiento resulte relevante por alguna razón particular, la elaboración del listado requerirá la aplicación de procedimientos adicionales. En este segundo caso, la primera fase consistiría en determinar cuáles son las unidades léxicas que, a juicio de los investigadores, deben conocer determinados sujetos. Puesto que tomar una decisión al respecto puede resultar complejo, se sugiere la aplicación de dos instrumentos a un grupo cuyo conocimiento respecto del área temática en estudio se estime como óptimo:

a) El primero es un test exploratorio de disponibilidad léxica en relación con el o los centros de interés de mayor pertinencia; la metodología de aplicación recomendada es la que respalda el Proyecto Panhispánico de Léxico Disponible (DISPOLEX, Grupo de Investigación, 2003-2013): se entrega a cada informante un set de hojas en blanco y un lápiz y se le pide anotar en columnas de arriba a abajo todas las palabras que vengan a su memoria en relación con cada uno de los centros de interés mencionado; para cada centro de interés, se asignan dos minutos.

b) El segundo es un cuestionario de aplicación individual, donde se solicita anotar en una hoja en blanco, también en un lapso de dos minutos, los vocablos que, a juicio de los informantes, son los más relevantes de acuerdo con los objetivos de creación del instrumento. Los datos obtenidos del test de disponibilidad léxica deben ser transliterados, estandarizados, lematizados y procesados estadísticamente, ya sea por medio de los programas Lexidisp o Dispogen, o del software Dispolex, usando la fórmula de López Chavez & Strassburger (1987, 1991), que considera la f_i de cada unidad léxica en cada rango y el número de sujetos que conforman la muestra.

Los datos obtenidos en ambos tests deben ser transliterados y estandarizados, para permitir el adecuado cotejo entre las respuestas de los distintos informantes, y organizarse de mayor

a menor disponibilidad, en el primer caso, y de mayor a menor $f_{i,}$ en el segundo. Los vocablos seleccionados de acuerdo con su relevancia estadística serán considerados vocablos meta y serán incluidos en el instrumento para medir si resultan o no familiares para los hablantes: conformarán el conjunto número 1.

3.1.2 La segunda fase consistirá en seleccionar dos conjuntos más, del mismo tamaño que el primero, de acuerdo con dos criterios: familiaridad y composición fonológica. El conjunto 2 corresponderá a vocablos altamente familiares y el 3 a vocablos medianamente familiares; si la investigación se centra en un área temática específica, lo ideal es que todos ellos se centren en las mismas áreas. Es necesario establecer estos dos conjuntos léxicos para poder compararlos con el conjunto 1, formado por los vocablos que se evaluarán como familiares o no familiares.

Para determinar qué unidades léxicas incluir en los conjuntos 2 y 3 se estimará su grado de familiaridad a partir de sus índices de U, f_i y D en el *Corpus Básico del Español de Chile* (2012) y de sus índices de disponibilidad léxica en obras disponibles para consulta, que idealmente describan la realidad lexicoestadística de la misma comunidad de habla que la investigación específica se propone estudiar.

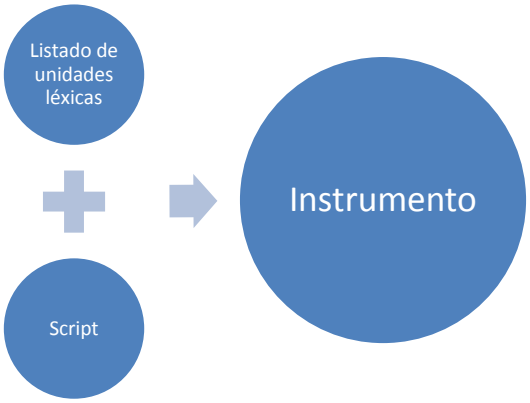
3.1.3 Se generará una equivalencia fonológica entre tríos de palabras pertenecientes a cada conjunto. Este paso es muy importante ya que, como se señaló en el marco teórico, las mediciones de duración se pueden ver afectadas por la composición fonológica segmental de la unidad léxica. Mientras más segmentos tenga la palabra y mientras más compleja sea su estructura silábica, su duración intrínseca será mayor, independientemente de la elevada o escasa familiaridad que tenga con ella el informante. Por ello, los vocablos seleccionados deberán pasar por un proceso de neutralización fonológica: los tres conjuntos estarán conformados por tríos lo más equivalentes posibles a nivel segmental. Esto quiere decir que un trío compuesto por un vocablo evaluado, uno familiar y uno moderadamente familiar, deben compartir las siguientes características: (1) número de sílabas, (2) tipo de estructura silábica, (3) posición de la sílaba tónica. Para elaborar estos tríos se tomará como punto de partida el conjunto evaluado y, a partir de este, se escogerán aquellas unidades léxicas que mejor cumplan con estos requisitos y que formen parte del proceso de selección lexicológica descrito en la fase previa. De esta manera, se neutraliza la influencia que pudiera tener la composición fonológica y la cantidad de segmentos que posee la unidad léxica, evitando que afecte las mediciones que se realizarán posteriormente a la lectura en voz alta de cada vocablo.

3.2. Presentación del instrumento

Por un lado, el instrumento consistirá en una lista de vocablos que el informante deberá leer en voz alta en un cierto orden y sin una cantidad de tiempo como máximo. La lista estará compuesta por los conjuntos de términos 1, 2 y 3 mencionados en la sección anterior, ordenados de manera aleatoria. El objetivo del instrumento será evaluar la familiaridad que tienen los hablantes con unidades léxicas estimadas como relevantes.

Por otro lado, el instrumento supondrá la creación de un script en Praat (Boersma & Weenik, 2014) que realizará el análisis y la interpretación que se indican en 3.4.

Figura 1. Composición del instrumento



3.3. Método de aplicación

Los sujetos tendrán que leer en voz alta la lista de palabras mencionada en el apartado anterior, la que se les entregará impresa, mientras su lectura es registrada por un investigador a través de una grabadora digital, individualmente y en una habitación con el menor ruido posible. El investigador les dará las instrucciones, señalando que deben leer la lista sin detenerse ni hacer preguntas, en el orden señalado en el listado.

Los informantes no deben tener problemas visuales ni de articulación vocal que puedan impedir o retrasar la lectura de las palabras de la lista. Quien realiza la grabación deberá determinar si la lectura se está viendo afectada por algún impedimento físico visual o vocal. Es conveniente tener en cuenta, a la hora de analizar los resultados, el nivel educativo y el perfil sociológico de cada informante, que eventualmente podría interferir en su capacidad de lectura.

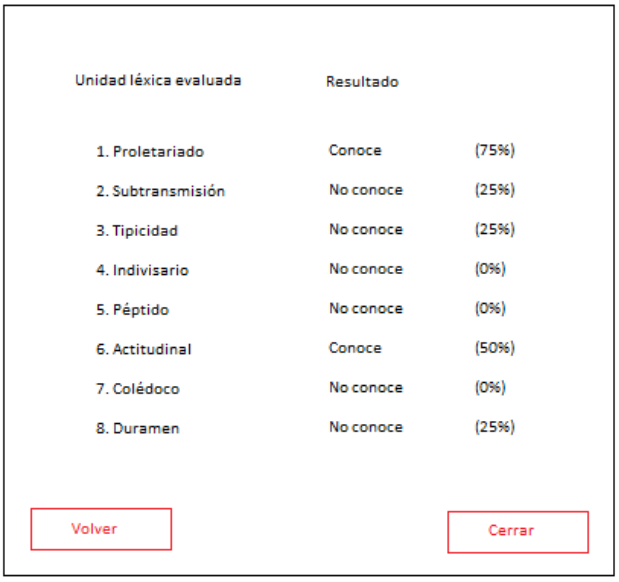
3.4. Método de análisis e interpretación

Las grabaciones realizadas a la lectura de los sujetos serán analizadas automáticamente por medio de un script programado en Praat (Boersma & Weenik, 2014). El programa medirá: (1) duración total de cada unidad léxica, (2) movimiento final de la curva melódica de cada unidad léxica, (3) tiempo que el hablante demora en comenzar la lectura de la unidad léxica siguiente (latencia).

En este trabajo, se prefirió el análisis suprasegmental por sobre el segmental, porque como se señaló en 2, si bien en inglés la inexacta pronunciación de una palabra puede ser indicadora de desconocimiento de la unidad léxica, esto se debe a las diferencias marcadas entre el sistema fonológico y el grafémico. En español, en cambio, esta divergencia no es tan acentuada, lo cual permite que, aun cuando un hablante de español desconozca una unidad léxica, pueda pronunciar de manera adecuada los segmentos que la componen.

Los resultados que arroje el script en esta primera instancia serán analizados estadísticamente para observar el comportamiento de los datos, establecer diferencias entre los distintos grupos y crear un criterio de evaluación que permita definir qué valores límites son pertinentes para determinar si el conjunto evaluado presentó más similitudes con el conjunto familiar o con el medianamente familiar, y así proponer una evaluación automática de los resultados en el script de Praat (Boersma & Weenik, 2014), modificándolo en lo que sea necesario. Es decir, el análisis estadístico permitirá establecer un modelo de evaluación para que en instancias posteriores el script no sólo analice de manera automática sino que también evalúe de manera automática las unidades léxicas del conjunto 1, entregando un resultado binario para cada una en términos de familiaridad o no familiaridad (conoce / no conoce), lo que se complementaría con información sobre el porcentaje de familiaridad (0%= el término es completamente desconocido / 100% = el término es altamente familiar"), como se muestra en la Figura 2.

Figura 2. Ejemplo de archivo output que podría entregar el instrumento



3.5. Comprobación de hipótesis

Se llevó a cabo un procedimiento piloto de aplicación del instrumento a 20 informantes: 10 académicos de la Facultad de Letras de la Pontificia Universidad Católica de Chile (PUC) y 10 estudiantes de las facultades de Letras y Educación de la PUC. Para elaborar la lista del piloto usamos los criterios señalados en la sección 3.1 para el establecimiento de los conjuntos 2 y 3; para la determinación de los vocablos meta, seleccionamos unidades léxicas de baja f_i y bajo U en el *Corpus Básico del Español de Chile* (cf. Figura 3).

Figura 3. Índices estadísticos de los vocablos de los conjuntos 1, 2 y 3

Vocablos de baja familiaridad (meta)	Vocablos mediana familiaridad	Vocablos alta familiaridad
•**Media Uso = 0,221** •U≤1,77 •F≤ 4	•**Media Uso = 27,693** •U ≥ 4,4 •F≥ 9	•**Media Uso= 133,59** •U ≥ 68,06 •$F_i \geq 83$

La lista finalmente consistió en los siguientes ítemes léxicos, ordenados por conjunto en la Figura 4.

Figura 4. Lista de unidades léxicas utilizadas en el piloto. El conjunto 2, ubicado en la primera columna, funciona como distractor; el conjunto 3, ubicado en la segunda columna, consta de palabras de mediana familiaridad y estructura semejante al conjunto de unidades meta, ubicado en la tercera columna.

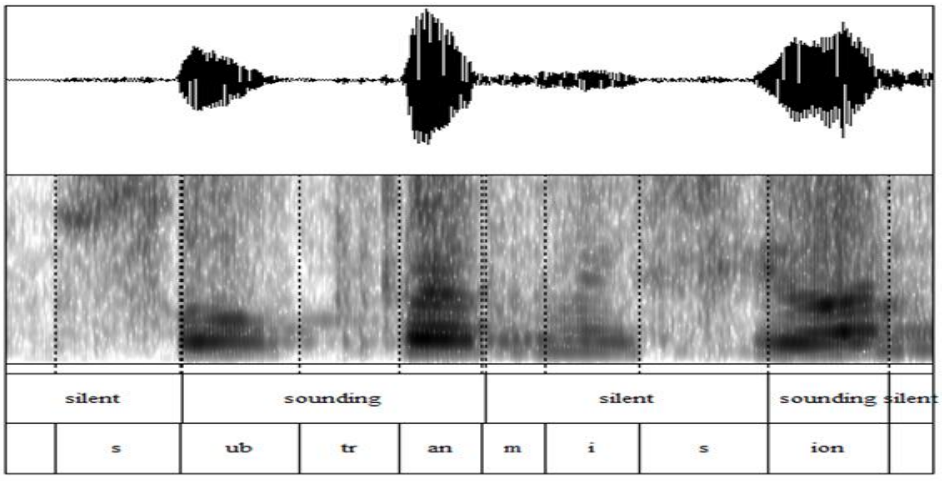
Conjunto 2: vocablos altamente familiares	**Conjunto 3: vocablos medianamente familiares**	**Conjunto 1:vocablos meta**
primera	protagonista	proletariado
subordinado	reconstrucción	subtransmisión
verdad	capacidad	tipicidad
aspecto	autoritario	indivisario
historia	pérdida	péptido
situación	intelectual	actitudinal
actual	católico	colédoco
transformar	dejaron	duramen

Dado que los vocablos raros tienden a presentar mayor complejidad de la estructura silábica y mayor número de sílabas, en muchos casos no es posible encontrar pares léxicos compuestos por unidades léxicas de alta y baja familiaridad; resulta, en cambio, más factible, encontrar unidades léxicas de similar complejidad de su estructura silábica y de similar número de sílabas. Por lo mismo, en esta etapa decidimos comparar los vocablos meta (raros) con unidades léxicas de mediana familiaridad. Por ello, la comparación se efectuó entre las unidades léxicas del conjunto 1 (en adelante, *raras*) y el conjunto 3 (en adelante, *familiares*). Las del conjunto 2, actuaron únicamente como distractores.

Las unidades léxicas se organizaron en una única columna, donde cada vocablo meta (conjunto 1, raras) iba precedido de sus dos pares: el primero, de alta familiaridad pero sin compartir estructura fonológica (conjunto 2, distractores); el segundo, de mediana familiaridad y mayor semejanza fonológica (conjunto 3, familiares).

Los sujetos fueron grabados, previa firma de un consentimiento informado, y sus grabaciones fueron procesadas a través del script en Praat ya mencionado. Sin embargo, no fue posible realizar una segmentación completamente automática, debido a la naturaleza de la grabaciones y el procedimiento que Praat utiliza para segmentar audios de forma automática: este programa trabaja con una función llamada "TextGrid Silences" la cual crea una planilla en que se marcan los segmentos con sonido y sin sonido en el audio. Para realizar este procedimiento, Praat opera con los siguientes parámetros: un límite de intensidad en decibeles, en que se señala el valor máximo de silencio con respecto al valor máximo de intensidad; una duración mínima de un intervalo de silencio; y la duración mínima de un intervalo de sonoridad (Boersma & Wernick, 2014). Algunos de los sujetos, al momento de la grabación, no dejaban una pausa notoria entre cada palabra, lo cual impedía que este la función pudiera delimitarlas con precisión, o también surgieron complicaciones debidas a la naturaleza fonética de las palabras como se demuestra en la Figura 5.

Figura 5. Segmentación automática de silencios en Praat: arriba se muestra el oscilograma de "subtransmisión"; más abajo aparece el espectrograma de la señal; a continuación se aprecia la planilla automática que genera Praat y bajo esta, por último, una planilla realizada a mano



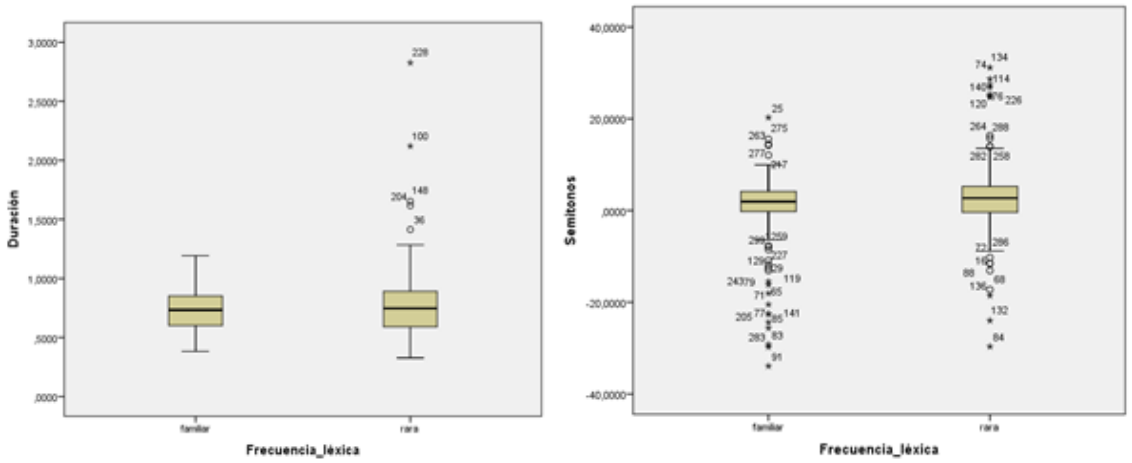
En este ejemplo, se observa cómo <subtransmisión> es incorrectamente segmentada por Praat, debido a la presencia interna de segmentos fricativos sordos, que no poseen frecuencia fundamental y se generan con poca intensidad. Debido a casos como este se prefirió, durante el piloto, segmentar manualmente los audios. Sin embargo, será necesario corregir y mejorar este mecanismo para en un futuro realizar una segmentación totalmente automática (cf. 5).

Si bien la segmentación fue manual, el procesamiento y recogida de datos a partir de las planillas se hizo de manera automática en base a scripts: para cada unidad léxica se obtuvieron los valores de duración, semitonos y latencia; en esta etapa, se observó la conveniencia de incorporar una cuarta variable suprasegmental: la intensidad. Los datos almacenados fueron procesados estadísticamente en el programa SPSS.

4. Resultados preliminares a partir del piloto

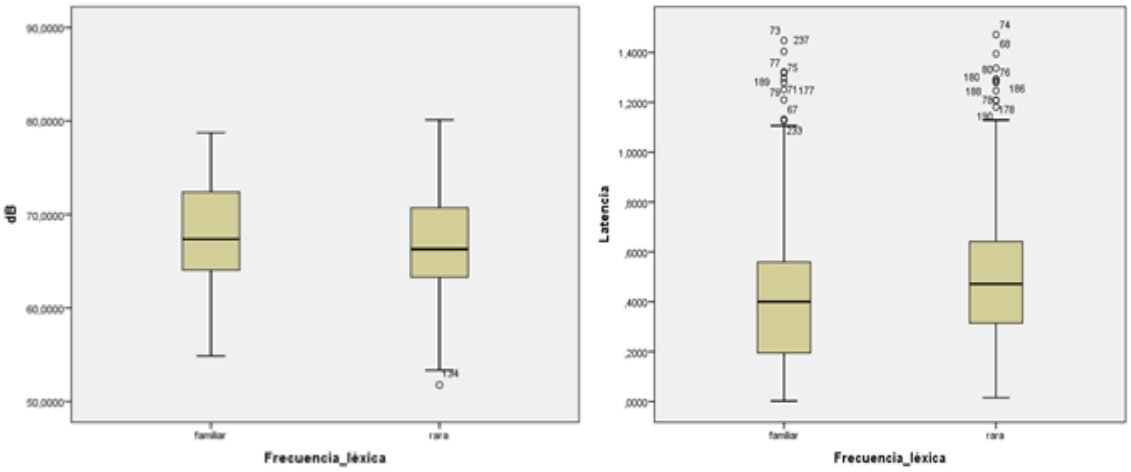
En un primer análisis exploratorio con el programa SPSS, se pudo observar la distribución general de las variables.

> Figura 6 y Figura 7. A la izquierda se muestra un gráfico de "cajas y bigotes" de la distribución de los valores de duración (eje vertical, en segundos) con respecto al grupo de palabras "familiares" y "raras" (eje horizontal). A la derecha, el gráfico muestra los valores de semitonos (eje vertical, en st) con respecto a los mismos dos grupos



En términos generales, la Figura 6 muestra que la duración se comportó de manera homogénea. Se puede ver una pequeña diferencia entre el grupo "familiar" y "rara", siendo las unidades léxicas de este último conjunto levemente más extensas que las primeras (recordar que fonológicamente las unidades léxicas de ambos conjuntos tienen la misma extensión). La Figura 7 indica que los semitonos presentan numerosos datos atípicos o marginales (que se salen del 95% de los casos). Los datos familiares presentan más datos atípicos en los semitonos negativos (entonación descendente), mientras que las palabras raras se extienden más hacia los semitonos positivos.

Figura 8 y Figura 9. A la izquierda se muestra un gráfico de "cajas y bigotes" de la distribución de los valores de intensidad (eje vertical, en decibeles) con respecto al grupo de palabras "familiares" y "raras" (eje horizontal). A la derecha, el gráfico muestra los valores de latencia (eje vertical, en segundos) con respecto a los mismos dos grupos



Como muestra la Figura 8, los valores de intensidad aumentan solo levemente en el caso de las palabras familiares. No hay mayores diferencias para esta variable comparando ambos grupos. En cuanto a la latencia (cf. Figura 9), sus valores aumentan para el grupo de palabras "raras".

Luego de este análisis exploratorio, se realizaron diferentes pruebas T agrupando distintas variables. Primero, se aplicó la prueba T comparando los grupos de unidades léxicas familiares y raras con respecto a las cuatro variables de nuestras hipótesis, usando como criterio un valor de significatividad del 5%. Los resultados se expresan en la Figura 10.

Figura 10. Valores p para las cuatro variables estudiadas al realizar la comparación entre el conjunto 3 (familiares) y el conjunto 1 (raras) de unidades léxicas. Las variables que presentan correlación significativa con la familiaridad se destacan en negritas

Variable	**Valor p**
Duración	0,089
Semitonos	**0,009**
Decibeles	0,214
Latencia	**0,027**

Los valores obtenidos para cada variable a través de esta prueba nos muestran que comparando ambos grupos totales de unidades léxicas (con todas las emisiones de todos los sujetos), solamente las variables de semitonos y latencia muestran una diferencia estadísticamente significativa. Para obtener resultados más específicos, luego se realizó la misma prueba T pero comparando los valores de cada par de palabras rara-familiar equivalente en su estructura fonológica. Los resultados se expresan en la Figura 11.
Sólo un par obtuvo diferencia significativa en dos variables (Par 8: <dejaron>-<duramen>). Hubo cuatro pares con diferencia significativa en una sola variable (Par 1: <protagonista>-<proletariado>, Par 2: <reconstrucción>-<subtransmisión>, Par 4: <autoritario>-<indivisario>, Par 6: <intelectual>-<actitudinal>). Sin embargo, también se presentaron pares que no obtuvieron ninguna diferencia significativa (Par 3: <capacidad>-<tipicidad>, Par 5: <pérdida>-<péptido> y Par 7: <católico>-<colédoco>). La entonación fue dos veces significativa, al igual que la duración; esta última, en el par 8, se comportó como altamente significativa. Es importante recordar que el tamaño de la muestra es reducido y, por lo tanto, la probabilidad de encontrar diferencias altamente significativas es baja. La interpretación de estos resultados se realizará a continuación.

Figura 11. Comparación de los resultados de cada variable para cada par compuesto por una palabra del conjunto 3 y otra del conjunto 1 que poseen equivalencia en el nivel de la estructura fonológica

	Duración	Semitonos	Decibeles	Latencia
Par 1	0,293	**0,025**	0,207	0,702
Par 2	**0,003**	0,854	0,886	0,370
Par 3	0,500	0,242	0,317	0,120
Par 4	0,114	0,558	0,228	**0,025**
Par 5	0,377	0,966	0,137	0,784
Par 6	0,064	**0,027**	0,575	0,360
Par 7	0,888	0,306	0,303	0,676
Par 8	**0,000**	0,928	**0,049**	0,669

5. DISCUSIÓN

Dado el carácter exploratorio y de piloto del presente estudio, los resultados no permiten comprobar ni rechazar de manera tajante las hipótesis planteadas. Sin embargo, ha sido una experiencia útil que será de ayuda al momento de llevar a cabo un trabajo más definitivo en esta área.

5.1 En primer lugar, el diseño del experimento y de la toma de muestras debe sufrir algunos cambios importantes que pueden haber tenido influencia a la hora de establecer la significatividad de los resultados:

5.1.1 En cuando a la selección de las unidades léxicas estimadas como "raras" para efectos del piloto, se consideró su frecuencia y uso de acuerdo con el *Corpus Básico del Español de Chile*. Si bien este criterio parece adecuado, debe ser perfeccionado:

a) Es posible que algunas de las unidades léxicas infrecuentes o de bajo uso pertenezcan por afinidad fonosemántica a una familia léxica mayor, lo que las haya vuelto familiares para los informantes; así, <subtransmisión> podría ser reconocida, pese a su bajo uso, por la mayor frecuencia de <transmisión> y <transmitir> y por el gran número de unidades léxicas que contienen el formante <sub> (cf. 2.1). Es preciso, entonces, considerar esta variable para efectos de la comprobación o refutación de las hipótesis respecto de la correlación entre familiaridad léxica y suprasegmentos.

b) Algunas unidades léxicas de bajo uso podrían ver aumentada su familiaridad a causa de su mayor disponibilidad léxica; este podría ser el caso de, por ejemplo, <proletariado>. Por ello, resulta relevante revisar el índice de disponibilidad léxica de las unidades pretendidamente "raras", para no obtener falsos negativos en la comprobación de las hipótesis.

c) En un nuevo piloto, sería conveniente aplicar un breve cuestionario semasiológico tras la aplicación del instrumento, para ratificar o descartar el grado de familiaridad de los informantes con las unidades léxicas estimadas como "raras".

d) Por último, para la aplicación definitiva del instrumento, una vez comprobadas o refutadas las hipótesis, es conveniente afinar aún más la selección de los vocablos meta. En el caso de que este instrumento se aplicara en dominios técnicos, como por ejemplo la salud o la ingeniería, se podrían complementar los métodos señalados en 3 con otros como el método Delphi, en que un grupo de expertos en el área de interés filtra los términos hasta llegar a un grado de acuerdo mínimo, o incorporar métodos del procesamiento del lenguaje natural y de corpus, como la extracción de términos técnicos desde corpus especializados.

5.1.2 Respecto del orden de las unidades léxicas en el listado:

a) La lista leída por los sujetos ordenaba las unidades léxicas de tal manera que los pares analizados se presentaban contiguos; es decir, <dejaron> y <duramen> se presentaban en orden sucesivo. Esto puede haber ayudado a los hablantes a pronunciar de manera más fluida la segunda unidad léxica del par habiendo leído justo antes la primera. Por lo tanto, es necesario aleatorizar la lista y posiblemente incluir más unidades léxicas como distractores. Los vocablos meta no deben ubicarse inmediatamente después de sus pares fonológicos, para que la lectura de los primeros no opere como un entrenamiento que facilite la posterior lectura de los vocablos meta por parte del informante.

b) Los vocablos meta y sus pares fonológicos no deben ubicarse ni en la primera ni en la última posición del listado, pues esto tiene implicancias en los suprasegmentos. Por

ejemplo, la última unidad léxica del listado tenderá a ser leída con entonación descendente, lo que anulará esa variable como predictora de familiaridad.

5.2 En segundo lugar, es necesario mejorar el procedimiento de análisis automático de Praat –el cual comete errores como el señalado en la sección 3.5– o considerar otros algoritmos que permitan realizar reconocimiento automático de regiones de habla y silencio en la señal, de tal manera que el procedimiento completo se pueda hacer de manera completamente automática. Este un objetivo muy relevante, considerando las posibles proyecciones de este proyecto (cf. 6).

5.3 En tercer lugar, en cuanto a los resultados, ninguna de las variables se comportó de manera constante en los ocho pares de unidades léxicas que constituían nuestro conjunto de observación. Por una parte, es necesario analizar el comportamiento de los pares aislados y realizarnos varias preguntas: ¿por qué hubo pares en que ninguna de las variables fue significativa?; podríamos suponer que dada la presentación de la lista de palabras, como se señaló recién, se facilitó la lectura de estos pares que, en el caso de los pares 3, 5 y 7 presentan una alta equivalencia en cuanto a su estructura fonológica; en el caso del Par 1, es posible que dado el bagaje cultural de los sujetos de la muestra <proletariado> haya resultado familiar a pesar de sus valores de uso y frecuencia; esta misma unidad léxica, por lo demás, pese a su baja frecuencia podría presentar una disponibilidad media que la hiciera familiar. Por otra parte, en términos globales, no necesariamente coinciden los valores significativos al comparar los conjuntos completos con los valores por cada par. Esto lógicamente se debe al tamaño de la muestra que en aplicaciones futuras del proyecto debe considerar una mayor.

5.4 Por último, de la mano con que el instrumento sea completamente automático, sería necesario crear una interfaz gráfica amigable con el usuario para que cualquier persona pueda aplicar este instrumento sin la necesidad de tener conocimientos técnicos en las áreas del léxico y la fonética, transformándolo en un software independiente de Praat y capaz de ser utilizado en otras plataformas que sean útiles para los usuarios.

6. PROYECCIONES Y CONCLUSIONES

Se proyecta aplicar este instrumento en ámbitos especializados como método de evaluación de la familiaridad léxica de los sujetos con las unidades léxicas. En este sentido se está desarrollando el instrumento en el área de la salud, para evaluar la familiaridad de pacientes crónicos cardiovasculares con unidades léxicas relevantes para el automanejo de su condición de salud (Proyecto de Investigación INTERDISCIPLINA Nº 1/2014 VRI UC (2014-2016)). En este proyecto, se busca establecer de manera rápida y eficiente si los pacientes con enfermedades crónicas cardiovasculares conocen o no unidades léxicas relevantes para el automanejo de su condición (por ejemplo, un paciente con diabetes debe conocer unidades léxicas como "insulina"), lo cual permitirá a profesionales de centros de atención primaria y secundaria evaluar rápidamente si el paciente está o no en condiciones de comprender las instrucciones específicas sobre su autocuidado, o si requiere explicaciones previas sobre términos relevantes para el automanejo de su condición que le son desconocidos.

En conclusión, si bien este proyecto aún está en una etapa exploratoria, desde una mirada aplicada se proyecta con aplicaciones útiles y relevantes para la sociedad y, desde el ámbito teórico, colabora a descubrir y estudiar las relaciones que existen entre el nivel léxico y el nivel fonético de la lengua. A la luz de los resultados obtenidos, descartamos la correlación entre familiaridad léxica e intensidad y continuaremos profundizando en la correlación entre dicha variable y las variables suprasegmentales entonación, latencia y duración.

Referencias bibliográficas

Ávila, Antonio M. 1998. *Elaboración, anotación y análisis del corpus oral del Proyecto V.U.M. Léxico de frecuencia del español hablado en la ciudad de Málaga.* Málaga: Universidad de Málaga.

Boersma, Paul & David Weenink. 2014. *Praat: doing phonetics by computer* [Programa computacional]. Version 5.3.77, retrieved 18 May 2014 from http://www.praat.org/

Castillo Fadić, María Natalia. 2001. *Los llamados extranjerismos en el diccionario de la Real Academia Española: criterios de selección y adaptación. Análisis metalexicográfico y reformulación*. Tesis de magíster en Lingüística. Santiago, Chile: Pontificia Universidad Católica de Chile.

Castillo Fadić, María Natalia. 2012a. *Corpus Básico del Español de Chile* ©.

Castillo Fadić, María Natalia. 2012b. *Léxico Básico del Español de Chile*. Tesis de doctorado en filología hispánica. Valladolid, España: Universidad de Valladolid.

Castillo Fadić, María Natalia. 2015. Léxico Básico del Español de Chile: el proyecto. *E-Aesla* 1. 1-8, http://cvc.cervantes.es/lengua/eaesla/pdf/01/51.pdf (10 de enero de 2016).

Cid Uribe, Miriam E., Héctor Ortiz-Lira, Mario Poblete Vallejos Hernán Pons Galea & José Luis Samaniego A. 2000. Hacia una descripción prosódica del español culto de Santiago de Chile: resultados de una investigación. *Onomázein* 5: 95-106.

Compton, Donald & Carlisle, Joanne. 1994. Speed of word recognition as a distinguishing characteristic of reading disabilities. *Educational Psychology Review* 6, 2: 115-140.

Davies, Robert, Javier Rodríguez-Ferreiro, Paz Suárez & Fernando Cuetos. 2013. Lexical and sub-lexical effects on accuracy, reaction time and response duration: impaired and typical word and pseudoword reading in a transparent orthography. *Read Writ* 26: 721-738.

DispoLex, Grupo de Investigación. 2003-2013. ¿Qué es el Proyecto Panhispánico? *Dispolex* http://www.dispolex.com/info/el-proyecto-panhispanico (17 de mayo de 2013)

Fernández Planas, Ana María. 2011. *Así se habla: nociones fundamentales de fonética general y española*. Barcelona: Horsori.

Gil, Juana. 2007. *Fonética para profesores de español: de la teoría a la práctica.* Madrid: Arco Libros.

López Chávez, Juan y Carlos Strassburger. 1987. Otro cálculo del índice de disponibilidad léxica. En *Actas del IV Simposio de la Asociación Mexicana de Lingüística Aplicada. Presente y perspectivas de la lingüística computacional en México*. México D.F.: UNAM.

López Chávez, Juan & Juan Carlos Strassburger Frías. 1991. Un modelo para el cálculo del índice de disponibilidad léxica individual: enseñanza del español como lengua materna. En Humberto López Morales (ed.), *Actas del II Seminario Internacional sobre Aportes de la Lingüística a la Enseñanza del Español como Lengua Materna*. 99-112. Río Piedras: Editorial de la Universidad de Puerto Rico.

Martínez, Eugenio & Ana María Fernández. 2007. *Manual de fonética Española.* Barcelona: Ariel.

McCormick, Christine & Jay Samuels. 1979. Word recognition by second graders: the unit of perception and interrelationships among accuracy, latency, and comprehension. *Journal of Reading Behavior* XI, 2: 107-118.

Muñoz-Builes, Diana, Dania Ramos, Domingo Román, Camilo Quezada, Héctor Ortiz-Lira, Magaly Ruiz y José Joaquín Atria. 2017. El habla ascendente de Chiloé: primera aproximación. *Onomázein* 37: 01-15.

Ortiz, Héctor, Marcela Fuentes, & Lluïsa Astruc. 2010. "Chilean Spanish Intonation". En Pilar Prieto & Paolo Roseano (eds), *LSPh 06: Transcription of Intonation of the Spanish Language.* Lincom Europa. 255-283.

Valencia, Sheila, Anthony Smith, Anne Reece, Min Li, Karen Wixson & Heater Newman. 2010. Oral reading fluency assessment: issues of construct, criterion and consequential validity. *Reading Research Quarterly* 45, 3: 270-291.

Wainschenker, Ruben, Jorge Doorn & Marcelo Castro. 2002. Quantitative values for perceptual notion of speech speed. *Medical Engineering & Physics* 24: 479-483.

Wright, Charles. 1979. Duration differences between rare and common words and their implications for the interpretation of word frequency effects. *Memory & Cognition* 7, 6: 411-419.