TESE DE DOUTORAMENTO

DEVELOPMENT AND PEDAGOGICAL APPLICATIONS OF AN AUDIO-TEXTUAL ENGLISH-SPANISH PARALLEL LITERARY CORPUS FOR THE STUDY OF ENGLISH PHONOLOGY

Michael Lang

ESCOLA DE DOUTORAMENTO INTERNACIONAL DA UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

PROGRAMA DE DOUTORAMENTO EN LINGÜÍSTICA

SANTIAGO DE COMPOSTELA

2021

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

DECLARACIÓN DEL AUTOR/A DE LA TESIS

D./Dña. **Michael Lang**

Título da tese: **Development and Pedagogical Applications of an Audio-Textual English-Spanish Parallel Literary Corpus for the Study of English Phonology**

Presento mi tesis, siguiendo el procedimiento adecuado al Reglamento y declaro que:

1) La tesis abarca los resultados de la elaboración de mi trabajo.
2) De ser el caso, en la tesis se hace referencia a las colaboraciones que tuvo este trabajo.
3) Confirmo que la tesis no incurre en ningún tipo de plagio de otros autores ni de trabajos presentados por mí para la obtención de otros títulos.

Y me comprometo a presentar el Compromiso Documental de Supervisión en el caso que el original no esté depositado en la Escuela.

En **Santiago de Compostela**, **a 29 de diciembre de 2020**.

Firma electrónica

AUTORIZACIÓN DEL DIRECTOR/TUTOR DE LA TESIS

D./Dña. **Irene Doval Reixa**

En condición de: Tutor/a y director/a

Título de la tesis: **DEVELOPMENT AND PEDAGOGICAL APPLICATIONS OF AN AUDIO-TEXTUAL ENGLISH-SPANISH PARALLEL LITERARY CORPUS FOR THE STUDY OF ENGLISH PHONOLOGY**

INFORMA:

Que la presente tesis, se corresponde con el trabajo realizado por D/Dña **Michael Lang**, bajo mi dirección/tutorización, y autorizo su presentación, considerando que reúne l os requisitos exigidos en el Reglamento de Estudios de Doctorado de la USC, y que como director/tutor de esta no incurre en las causas de abstención establecidas en la Ley 40/2015.

En **Santiago de Compostela**, **31 de diciembre de 2020**

Firma electrónica

AUTORIZACIÓN DEL DIRECTOR/TUTOR DE LA TESIS

D./ Dña. Xavier Gómez Guinovart

En condición de: Director

Título de la tesis: Development and Pedagogical Applications of an Audio-Textual English-Spanish Parallel Literary Corpus for the Study of English Phonology

INFORMA:

Que la presente tesis, se corresponde con el trabajo realizado por D. Michael Lang, bajo mi dirección/tutorización, y autorizo su presentación, considerando que reúne l os requisitos exigidos en el Reglamento de Estudios de Doctorado de la USC, y que como director/tutor de esta no incurre en las causas de abstención establecidas en la Ley 40/2015.

En Vigo, 30/12/2020

Firma electrónica

ACKNOWLEDGEMENTS

The work here would not have been made possible without all the support provided to me by many incredible people. It is with my deepest gratitude that I dedicate this doctoral dissertation to the following:

To **Xavier Gómez Guinovart**, my supervisor who gave me the initial push that started what has become a successful four-year journey. You've helped me take a simple idea and turn it into what has become a fantastic linguistic resource. Your skills and knowledge have made the corpus what it is. I could not have done this without you. ¡Viva LITTERA!

To **Irene Doval Reixa**, my supervisor, whose generous support and guidance has always kept me on track. Thank you for sharing your expertise with me at every turn and encouraging me to do the same with others. The workshops would not have been possible without you and provided me with invaluable experience.

To my family—**Mom, Dad, Ricky, Erik & Laurie**—whose unending love and encouragement is always with me, wherever I may be. Thank you for always believing in me. Next up…a J.O.B. (you can breathe easy now, mom).

To **Chaves & María**. I will never forget your friendship.

Finally, to **Irene**, who has pulled me through my moments of self-doubt and showed me what I am capable of. Thank you for listening with patience to all my digressions on linguistics and languages, and for being a source of continuous love and joy in my life.

ABSTRACT

The field of Data-Driven Learning (DDL) — an approach to second language learning in which the student interacts directly with corpus data — has made much progress in only the matter of a few decades. However, there are still certain frontiers that have thus far remained underexplored, mostly the result of limited technological capabilities for a good portion of the field's existence. Until now, DDL has mainly centered on text corpora, leaving aside such aspects of language learning as oral comprehension and speech production. This doctoral dissertation presents the LITTERA corpus, and examines in depth how this English-Spanish parallel literary speech corpus can be applied to language learning within the framework of DDL. The dissertation begins with a general overview of the current state of DDL, followed by a detailed description of the creation and design of the LITTERA corpus. Then a series of potential pedagogical exercises are presented, aimed at showing how LITTERA can be applied to the learning of English phonology by Spanish-speaking students. The exercises set out to examine how the different features of English prosody—co-articulatory phenomena such as linking, blending, assimilation, elision, resyllabification, palatization, as well as vowel reduction—can be studied in the data to improve students' oral comprehension and speech production. Furthermore, possible DDL question prompts are proposed to explore the different features in the classroom.

Keywords: Data-Driven Learning (DDL), speech corpora, English phonology, English as a Second Language (ESL)

TABLE OF CONTENTS

RESUMEN EN ESPAÑOL

INTRODUCCIÓN

Vivimos en una época en la que, gracias a la ubicuidad y la expansión constante de Internet, y de los aparatos *inteligentes* que utilizamos para acceder a ella, el aprendizaje de idiomas se realiza cada vez más en el universo digital a través de apps, redes sociales, servicios de *streaming* (Netflix, YouTube…) y plataformas digitales diseñadas para el aprendizaje de idiomas. Cada generación se familiariza más y es más hábil con las tecnologías que la anterior, y cada vez hay más estudiantes que parecen preferir los recursos digitales a los libros de texto y las metodologías tradicionales.

En las últimas décadas, se ha despertado un interés en trasladar las tecnologías lingüísticas—en concreto, los corpus en línea y los programas de concordancias—desde el ámbito de los investigadores expertos al de los estudiantes de idiomas no expertos. Las consecuencias de este desplazamiento se aprecian en el crecimiento del campo de estudio del *Data-Driven Learning* (DDL, o Aprendizaje Dirigido por los Datos). El DDL es una aproximación al aprendizaje de idiomas en la cual el estudiante interactúa directamente con los datos del corpus, en lugar de hacerlo indirectamente a través de medios convencionales como los libros de texto y los diccionarios. Hasta ahora, la gran mayoría de los estudios en DDL han trabajado con corpus de texto, a pesar de que hay cada vez más corpus multimedia en los que se puede encontrar también audio y vídeo. Por ello, hemos creado el corpus LITTERA, un corpus paralelo literario audiotextual inglés-español para el estudio de la fonología inglesa en el marco del DDL.

De este modo, el objetivo de esta tesis doctoral es doble: 1) por un lado, presentar y describir el corpus LITTERA en detalle, desde su concepción y creación hasta su composición y características actuales; 2) por otro lado, examinar en qué manera la utilización del corpus en el ámbito del DDL puede facilitar el aprendizaje de algunos aspectos de la fonología inglesa, en particular, a los hispanohablantes. Las aplicaciones didácticas elaboradas en el presente trabajo se basan en los datos reales del corpus. Es el deseo del autor que este trabajo dé lugar a futuras investigaciones empíricas que examinen la eficacia del corpus LITTERA como un recurso de DDL y que analicen su recepción y utilización por parte de los estudiantes universitarios de inglés.

La tesis está dividida en tres secciones. La primera presenta el concepto de DDL junto con un resumen del estado actual del campo. La segunda sección detalla la creación y las características del corpus LITTERA, y la tercera elabora una serie de aplicaciones didácticas para los diferentes fenómenos de la fonología inglesa que se pueden estudiar directamente en el corpus.

DATA-DRIVEN LEARNING

Se ha demostrado que la aproximación del DDL tiene muchos efectos positivos sobre el aprendizaje de idiomas, tal como una mejoría en vocabulario, en el reconocimiento de patrones léxico-gramaticales y en la interacción constante con los ítems de alta frecuencia. Al interactuar directamente con los datos del corpus, los estudiantes toman un papel activo en su aprendizaje, lo cual da lugar a un crecimiento de su autonomía, que contrasta con la forma tradicional pasiva en la que un profesor "deposita" información en la cabeza del estudiante. Esta participación activa promueve el "aprendizaje por descubrimiento" (*discovery learning*), en el que el estudiante forma sus propias reglas con la intervención mínima y necesaria del profesor para evitar conclusiones erróneas sobre los datos.

No obstante, el DDL también ha generado críticas, que suelen centrarse en tres aspectos diferentes: la tecnología, el conocimiento de los profesores y el mismo uso de los corpus. En cuanto a la tecnología, la cuestión de acceso supone un problema para los que carecen de una conexión a Internet, sea en el aula o en casa. Este obstáculo se puede remediar con la creación de actividades basadas en papel, como ya se ha hecho en algunos estudios. En los casos en los que la conexión no supone un problema, es necesario que el profesor tenga los conocimientos necesarios para introducir los corpus en la lección. Por desgracia, los profesores que están familiarizados con los corpus son la excepción, no la regla. Hay que tener en cuenta el hecho de que hoy en día muchos profesores ya están bajo mucha presión con la cantidad de contenido que tienen que impartir, lo cual supone otra posible barrera para que los corpus lleguen a las aulas. En los casos en los que sí entran en las aulas, es posible que surjan más problemas a la hora de realizar actividades. El lenguaje del corpus, aunque auténtico (en el sentido de que son ejemplos de uso real por nativos), puede resultar demasiado difícil para los estudiantes con niveles más bajos, y, en consecuencia, reducir el interés en la actividad. También es posible que los estudiantes se sientan abrumados por la cantidad de información que puede generar una simple búsqueda en el corpus. Por eso, es fundamental que, en la mayoría de los casos, los profesores perfilen las actividades a medida de las necesidades de los estudiantes. Otra solución se encuentra en el diseño del corpus a través de algunas técnicas como la anotación pedagógica o la posibilidad de limitar los resultados de la búsqueda. La involucración del profesor y el buen diseño del corpus no son una panacea, pero ayudan a minimizar las dificultades que pueda tener el usuario.

La recepción del DDL por parte del alumnado ha sido bien documentada en los últimos años gracias a que el número de los estudios empíricos ha ido creciendo. En el lado positivo, los estudiantes han reportado que disfrutan al trabajar con lenguaje auténtico y les gusta que se pueda ver el contexto del ítem en cuestión. Han usado los corpus como herramienta de referencia para comprobar y confirmar su propio uso de la lengua en cuestión. Por otro lado, además de las críticas citadas anteriormente, algunos estudiantes han comentado que les parece muy tedioso examinar muchas líneas de concordancia y que les resultaba difícil formular los términos de búsqueda adecuados.

La recepción por parte de los profesores ha sido positiva pero con un cierto grado de escepticismo. Las preocupaciones suelen estar relacionadas con el tiempo necesario para crear ejercicios de DDL y con su propia percepción como figuras de autoridad y conocimiento en el aula. Muchos tiene un punto de vista pragmático que gira entorno de la cuestión: *¿qué provecho puedo sacar de esto, dadas todas las presiones ya existentes?* Por tanto, no debería sorprender que los profesores puedan sentirse reacios con respecto a la idea de trabajar con corpus en el aula.

Por ello, el concepto de *training* (entrenamiento) es muy importante. No se puede esperar que un profesor o un estudiante sepa manejar un corpus sin haber recibido la formación necesaria. Eso no significa que esta formación tenga que ser muy extensa, sobre todo si el corpus tiene una interfaz intuitiva. El corpus LITTERA cuenta con una serie de ejercicios de introducción para que el usuario se pueda familiarizar con los textos del corpus y, con su composición, características y opciones de búsqueda. Los ejercicios forman parte de una serie de vídeo-tutoriales en YouTube creados por el autor. El enlace a los tutoriales se encuentra en la página principal del corpus LITTERA. En menos de media hora, habrán trabajado las habilidades necesarias para poder realizar sus propias búsquedas en el corpus.

EL CORPUS LITTERA

Muy pocos estudios de DDL se han enfocado en el tema de la fonología. Esto se debe en parte a que, durante muchos años, no había los recursos necesarios para crear un corpus con audio a la escala necesaria para tal estudio. Las limitaciones tecnológicas de las últimas décadas del siglo XX, junto con la falta de los recursos humanos necesarios, no permitían que este campo floreciera.

Para solucionar esta falta de corpus con audio (*speech corpora*) necesarios en el campo de DDL, hemos creado el corpus LITTERA, un corpus paralelo literario audiotextual inglés-español diseñado para el estudio de diferentes aspectos de la fonología inglesa en el marco de la enseñanza del ESL (*English as a Second Language*). El corpus está compuesto por 25 textos literarios en lengua inglesa y sus traducciones al castellano, así como segmentos de audio de los audiolibros correspondientes en inglés. Cuenta con casi dos millones de palabras (983.618 en inglés y 985.058 en español) y 63.508 unidades de traducción. Por cada unidad de traducción, hay también un segmento de audio que corresponde al texto en inglés. Por tanto, hay 63.508 archivos de audio. Su nombre se debe al dominio literario del corpus, ya que *littera* es la palabra latina de *letra*, y de ahí, *literatura*.

Los textos han sido seleccionados en gran parte en base a las obras y los autores incluidos en el Plan de Estudios del grado de Filología Inglesa de la Universidad de Santiago de Compostela. En la selección también ha influido la disponibilidad del texto original, la traducción y el audiolibro (en plataformas de uso abierto como YouTube o LibriVox). Libros pertenecientes a los géneros de literatura infantil y juvenil (como *The Hungry Hungry Caterpillar* y *Harry Potter and the Philosopher's Stone*, respectivamente) fueron añadidos para que hubiese más variedad de registros representada en el corpus. El autor ha sido responsable de la selección y alineación de los textos, y la segmentación de los audiolibros. También ha aportado ideas a las opciones disponibles en el diseño, aunque el trabajo técnico fue realizado por su co-director de tesis, Xavier Gómez Guinovart.

LITTERA forma parte de dos colecciones de corpus, el corpus CLUVI y el corpus SensoGal, los dos alojados en el Seminario de Lingüística Informática de la Universidad de Vigo. El corpus CLUVI facilita las consultas flexibles con expresiones regulares, mientras que los corpus compilados en el SensoGal están etiquetados semánticamente y permiten las búsquedas por forma (palabra o lema) o concepto (según los conceptos establecidos en la red semántica de WordNet y en su adaptación en Galnet). La interfaz de SensoGal también permite realizar consultas especificando la categoría gramatical. Las dos colecciones ofrecen opciones para controlar el número de resultados, elegir la variedad oral del inglés (americano o británico), ampliar el contexto, y, en el caso de LITTERA, reproducir el audio de los resultados.

El corpus LITTERA está pensado para estudiantes universitarios de habla hispana, aunque cualquier persona que aprenda inglés se puede beneficiar de él. Puesto que los textos fueron seleccionados a partir del programa de Filología Inglesa, la mayoría de las obras son relevantes para los estudios universitarios. Además, es posible que la disponibilidad del texto original con la traducción, junto con la posibilidad de buscar dentro de un solo texto si se desea, anime al estudiante a leer textos que antes no habría intentado leer por su dificultad.

LAS APLICACIONES DIDÁCTICAS

Las aplicaciones didácticas desarrolladas en la tesis se centran en diversos aspectos suprasegmentales de la fonología inglesa, concretamente, en diferentes características de la prosodia, tales como el *linking* (cómo las palabras se enlazan en el discurso fluido), la *asimilación* (cuando un fonema asimila las características de los fonemas alrededor), la *elisión* (la omisión de un fonema), la *palatización* (cuando un fonema frontal, e.g. /t/ o /d/, ocurre ante un fonema velar, e.g. /j/, y se produce un fonema africado como /tʃ/ o /dʒ/), las *formas reducidas* (el hecho de que muchas palabras gramaticales en inglés tienen una forma de citación, i.e. de diccionario, y una forma reducida para destacar las palabras que llevan más peso léxico) y la *unidad de entonación*. Aunque el audio no proviene de conversaciones espontáneas, los aspectos de la prosodia que se examinan aquí se encuentran en todos los discursos del inglés hablado. Por tanto, no es necesario que sea lenguaje espontáneo. Sin embargo, los audiolibros tienen la ventaja de que las obras contienen representaciones de diálogo que reflejan el lenguaje conversacional, lo cual, finalmente, permite un análisis de algunas de las características del inglés conversacional también.

El primer aspecto de la prosodia que se examina es el más general, las unidades de entonación (*tone units*), que son unidades prosódicas con al menos una sílaba, que es la tónica de la unidad, y que acaban con una entonación de frontera (la subida o caída de la entonación). Las unidades de entonación están separadas frecuentemente por signos de puntuación, aunque no en todos los casos. Poder reconocer donde se ubican las fronteras entre una unidad prosódica y otra es fundamental para luego entender por qué ciertos fonemas se realizan de diferentes formas. Las unidades de entonación se pueden examinar con cualquier búsqueda, ya que cada unidad de traducción contiene al menos una unidad de entonación.

El siguiente aspecto que se analiza es el concepto de *linking* (enlazamiento), es decir, cómo se conectan los fonemas a través de las fronteras entre palabras en el habla. En concreto, se examinan las combinaciones de C+V (cuando una palabra acaba en consonante y la siguiente empieza por una vocal) y C+C. Se puede encontrar muchos ejemplos en el corpus para mostrar los diferentes fenómenos a la hora de enlazar las palabras en el habla fluida, tales como la elisión, la asimilación o la resilabificación, que ocurre cuando un consonante en la coda de una sílaba se convierte en el ataque de la sílaba siguiente. Esto se analiza a través de los *phrasal verbs* en el corpus, como por ejemplo, el fonema /k/ siendo resilabificado en *take off*, lo cual puede causar la malinterpretación de *to cough*. Después se presenta el concepto de *blending* para analizar los ejemplos de C+C cuando las dos consonantes son la misma.

Una vez establecidas las características básicas del enlazamiento de palabras, se examina lo que los datos muestran sobre las combinaciones complejas de consonantes a través de las fronteras de palabras. El español no permite combinaciones complejas de consonantes al final de una palabra, lo cual genera muchas dificultades para los hispanohablantes al encontrarse con palabras como *asked* /æskt/. Más dudas sobre su realización pueden surgir al tener que combinar una palabra ya complicada por sí con otra palabra que empieza por un consonante. Los datos del corpus muestran los casos en los que se puede reducir la combinación de consonantes a través de la elisión de sonidos. Esta información es imprescindible no solo para la percepción de estos sonidos, sino también para facilitar su producción. Es importante que los estudiantes sean conscientes de cómo pueden simplificar las combinaciones complejas.

La cuarta aplicación didáctica trata la asimilación a través de las fronteras de palabras. Los datos del corpus contienen muchos ejemplos de este fenómeno, que se puede encontrar con facilidad gracias al uso de expresiones regulares. Un ejemplo es la asimilación de /v/ al final de *have* ante una consonante sorda, como en el caso de *have to*, convirtiéndola en /f/. Esto se compara con los casos en los que *have* es seguida por una vocal, como en el caso de *has a*, lo cual produce una consonante sonora /v/. Las expresiones regulares, por tanto, permiten buscar las dos formas en la misma búsqueda, lo cual sirve para compararlas y escuchar las diferencias sin tener que realizar búsquedas separadas. Otro ejemplo de asimilación es la palatización, que se puede apreciar en los datos del corpus, sobre todo en ítems de alta frecuencia como *what you* y *but you*, junto con otros ejemplos en los que un narrador elige hacerla con algunos personajes y no con otros, dando lugar a una variación estilística motivada en el discurso.

Con el análisis de los fenómenos fundamentales de la prosodia—asimilación y elisión de sonidos—, ya existe una base para examinar la realización de un morfema específico, el del pasado simple, *-ed*, como en *work**ed***. Aunque la pronunciación de este morfema, /t, d, ɪd/, se suele enseñar según la sonorización del fonema anterior (en los casos de /t/ y /d/) o el punto de articulación (en el caso de /ɪd/), los datos del corpus demuestran que hay otros factores que contribuyen a su realización, como los fonemas posteriores y la presencia de una frontera de una unidad de entonación. Por eso, es frecuente que los estudiantes no perciban la realización de este morfema debido a estos otros factores prosódicos, como la elisión y la asimilación. Esto puede causar una transformación inesperada en el morfema que no permite su percepción si uno solo espera escuchar los fonemas /t, d, ɪd/ de la regla original. Por tanto, en esta sección se proponen una serie de preguntas basadas en el DDL para encaminar el estudiante hacia el descubrimiento de los datos reales que puedan informar su propia comprensión y producción de este morfema.

A continuación, se explora la reducción vocálica en los datos del corpus. Uno de los ejemplos principales es el del verbo auxiliar *can*. Es común que los estudiantes se quejen de que muchas palabras parecen ‘desaparecer’ en el habla fluida en inglés. Esto se debe a que muchas palabras funcionales/gramaticales tienen al menos dos formas, la entera (i.e. de diccionario) y la reducida (influenciada por las características de la prosodia). Los datos del corpus muestran las dos formas y cómo *can* se puede reducir y por qué *can’t*, en cambio, nunca se reduce. También se puede apreciar ciertos casos en los que *can* no se reduce para mostrar énfasis o contraste. Esta distinción entre las dos formas es fundamental, ya que al usar la forma entera de *can* con la vocal abierta, es posible crear confusión para el interlocutor y que interprete *can* como *can’t*, sobre todo porque la elisión del fonema /t/ en posición final es algo frecuente (y que se puede apreciar en los datos). Por tanto, los estudiantes deberían ser conscientes de las dos formas para poder comprender un enunciado con *can* adecuadamente, y saber utilizarlo para no crear confusión con *can't*. También se examinan las formas reducidas del verbo auxiliar *have*. También se presenta aquí el concepto de clitización (por ejemplo, adjuntar *have* a otro verbo auxiliar, como *should* para producir una forma clitizada de *have*: *should've* o *shoulda*).

Dado que el corpus está compuesto por textos literarios, la siguiente aplicación didáctica trata sobre la representación del lenguaje hablado en el diálogo de los textos. Aunque LITTERA no contiene audio de conversaciones espontáneas, el diálogo escrito sirve como un sustituto de ello. A pesar de que los autores no suelen ser lingüistas profesionales, en muchos casos intentan captar la forma 'auténtica' o 'realista' en la que hablan las personas. Los autores sí son hablantes del idioma y quieren reflejar esos matices del habla conversacional que puedan servir para informar al lector sobre el personaje en cuestión. Por tanto, muchas características del inglés hablado y conversacional aparecen en los datos y señalan información importante sobre los efectos de la prosodia. Esto incluye representaciones escritas de características conversacionales tales como la clitización (lo cual incluye contracciones) y la reducción de palabras (por reducción vocálica o elisión). Algunos ejemplos de estos fenómenos representados en el diálogo escrito incluyen las contracciones de los verbos modales (e.g. *she's gone*, *there'll*, *what'll*, *it'd*), las formas clitizadas de *have* (*kinda*, *sorta*) y *to* (*gotta*, *gonna*, *wanna*), la elisión de sonidos al inicio de la palabra (*'em*, *'ud*, *'im*, *'as*, *'is*, y *'em*), y la reducción de *for* a una vocal media en *fer*. Todos estos ejemplos reflejan tendencias reales del inglés hablado. Esta información es extremadamente útil para la comprensión y producción natural del idioma.

Por último, se analiza un aspecto característico del inglés norteamericano: el *flap* (también conocido como el *tap*, o 'toque'). El *flap* ocurre cuando un consonante coronal /t/ o /d/ aparece en una posición intervocálica. Ocurre principalmente en el inglés norteamericano, aunque se puede encontrar en otras variedades de forma menos extensa. El *flap* se puede encontrar dentro de la palabra o entre dos palabras, siendo el segundo caso el más complicado para los estudiantes, ya que es menos predecible. Es más, el *flap* puede crear homófonos que en otras variedades no lo son, como es el caso de *seeded* y *seated*. A través de la opción de buscar en el corpus por variedad hablada (*Spoken English Variety*), se puede apreciar estas diferencias en la pronunciación. Aún es más complicado cuando el *flap* ocurre entre varias palabras seguidas, e.g. *let it out* y *put it out*. Esta sección demuestra cómo el profesor puede explorar este fenómeno en el corpus con los estudiantes, sobre todo con los ítems de alta frecuencia como *that* + V (e.g. *that everything*) o *get* + V (e.g. *get up*), para que sean conscientes de este fenómeno en ciertas variedades del inglés.

COMENTARIOS FINALES

El presente trabajo describe la creación y el diseño del corpus LITTERA y presenta algunas aplicaciones pedagógicas para el estudio de la fonología inglesa desde la perspectiva teórica de DDL en el ámbito de la enseñanza del ESL. El corpus LITTERA es un corpus paralelo literario audiotextual inglés-español pensado para estudiantes universitarios hispanohablantes de filología inglesa, ya que los textos literarios, en gran parte, fueron seleccionados en base al plan de estudios para ese grado en la Universidad de Santiago de Compostela.

El DDL es una aproximación al aprendizaje de idiomas en la que los datos del corpus sirven como *input* y substituyen a la forma tradicional de recibir la información pasivamente del docente. De esta forma, los estudiantes asumen un papel activo en su aprendizaje, lo cual conlleva una serie de ventajas, tanto a nivel lingüístico como personal. Puesto que la mayoría de los trabajos de DDL se centran en los corpus textuales, el autor tiene la esperanza de que esta tesis doctoral ayude a acercar los estudios de DDL a los corpus con audio (*speech corpora*).

Este trabajo muestra cómo un corpus con segmentos de audio puede servir como herramienta para el aprendizaje de la fonología inglesa en un contexto de ESL. Se analizan diferentes aspectos suprasegmentales del inglés que no se suelen estudiar demasiado con las metodologías tradicionales. De esta manera, los estudiantes, con la ayuda del docente, pueden hacer sus propias observaciones sobre los aspectos prosódicos del habla fluida, como la reducción vocálica, la elisión de sonidos, la asimilación, la clitización y la palatización, entre otras tendencias. De esas observaciones se pueden llegar a formar reglas que ayudarán en la comprensión oral y en la producción.

El autor espera que esta investigación dé lugar a estudios empíricos que permitan comprobar la eficacia de este corpus, o cualquier otro con audio, en el aprendizaje de la fonología en el aula. Las aplicaciones didácticas elaboradas aquí tienen como objetivo demostrar lo que dicen los datos del corpus y por qué vale la pena que los estudiantes trabajen con ellos. A pesar de que el campo de estudio del DDL ha crecido mucho en los últimos años, sigue habiendo mucho territorio sin explorar, del cual esta tesis doctoral ha intentado alumbrar una parte.

RESUMO EN GALEGO

INTRODUCCIÓN

Vivimos nunha época na que, grazas á ubicuidade e a expansión constante da internet, e dos aparellos *intelixentes* que utilizamos para acceder a ela, a aprendizaxe de idiomas realízase cada vez máis no universo dixital a través de apps, redes sociais, servizos de *streaming* (Netflix, Youtube…) e plataformas dixitais deseñadas para a aprendizaxe de idiomas. Cada xeración familiarízase máis e é máis hábil coas tecnoloxías que a anterior, e cada vez hai máis estudantes que parecen preferir os recursos dixitais aos libros de texto e as metodoloxías tradicionais.

Nas últimas décadas, espertouse un interese en trasladar as tecnoloxías lingüísticas—en concreto, os corpus en liña e os programas de concordancias—desde o ámbito dos investigadores expertos ao dos estudantes de idiomas non expertos. As consecuencias deste desprazamento apréciase no crecemento do campo de estudo do *Data-Driven Learning* (DDL, ou Aprendizaxe Dirixida polos Datos). O DDL é unha aproximación á aprendizaxe de idiomas na cal o estudante interactúa directamente cos datos do corpus, en lugar de facelo indirectamente a través de medios convencionais como os libros de texto e os dicionarios. Ata o de agora, a gran maioría dos estudos en DDL traballaron con corpus de texto, ainda que hai cada vez máis corpus multimedia nos que se pode atopar tamén audio e mais vídeo. Por iso, creamos o corpus LITTERA, un corpus paralelo literario audiotextual inglés-español para o estudo da fonoloxía inglesa no marco do DDL.

Deste xeito, o obxectivo desta tese doutoral é dobre: 1) por unha banda, presentar e describir o corpus LITTERA en detalle, desde a súa concepción e creación ata a súa composición e características actuais; 2) doutra banda, examinar en que maneira a utilización do corpus no ámbito do DDL pode facilitar a aprendizaxe dalgúns aspectos da fonoloxía inglesa, en particular, aos hispanofalantes. As aplicacións didácticas elaboradas no presente traballo baséanse nos datos reais do corpus. É o desexo do autor que este traballo dea lugar a futuras investigacións empíricas que examinen a eficacia do corpus LITTERA como un recurso de DDL e que analicen a súa recepción e utilización por parte dos estudantes universitarios de inglés.

A tese está dividida en tres seccións. A primeira presenta o concepto de DDL xunto cun resumo do estado actual do campo. A segunda sección detalla a creación e as características do corpus LITTERA, e a terceira elabora unha serie de aplicacións didácticas para os diferentes fenómenos da fonoloxía inglesa que se poden estudar directamente no corpus.

DATA-DRIVEN LEARNING

Demostrouse que a aproximación do DDL ten moitos efectos positivos sobre a aprendizaxe de idiomas, tal como unha mellora no vocabulario, no recoñecemento de patróns léxico-gramaticais e na interacción constante cos ítems de alta frecuencia. Ao interactuar directamente cos datos do corpus, os estudantes toman un papel activo na súa aprendizaxe, o cal dá lugar a un crecemento da súa autonomía, que contrasta coa forma tradicional pasiva na que un profesor "deposita" información na cabeza do estudante. Esta participación activa promove a "aprendizaxe por descubrimento" (*discovery learning*), no que o estudante forma as súas propias regras coa intervención mínima e necesaria do profesor para evitar conclusións erróneas sobre os datos.

Con todo, o DDL tamén xerou críticas, que adoitan centrarse en tres aspectos diferentes: a tecnoloxía, o coñecemento dos profesores e o mesmo uso dos corpus. En canto á tecnoloxía, a cuestión de acceso supón un problema para os que carecen dunha conexión a Internet, sexa na aula ou en casa. Este obstáculo pódese remediar coa creación de actividades baseadas en papel, como xa se fixo nalgúns estudos. Nos casos nos que a conexión non supón un problema, é necesario que o profesor teña os coñecementos necesarios para introducir os corpus na clase. Por desgraza, os profesores que están familiarizados cos corpus son a excepción, non a regra. Hai que ter en conta o feito de que hoxe en día moitos profesores xa están baixo moita presión coa cantidade de contido que han impartir, o cal supón outra posible barreira para que os corpus cheguen ás aulas. Nos casos nos que si entran nas aulas, é posible que xurdan máis problemas á hora de realizar actividades. A linguaxe do corpus, aínda que auténtico (no sentido de que son exemplos de uso real por nativos), pode resultar demasiado difícil para os estudantes con niveis máis baixos, e, en consecuencia, reducir o interese na actividade. Tamén é posible que os estudantes se sentan atafegados pola cantidade de información que pode xerar unha simple procura no corpus. Por iso, é fundamental que, na maioría dos casos, os profesores perfilen as actividades a medida das necesidades dos estudantes. Outra solución atópase no deseño do corpus a través dalgunhas técnicas como a anotación pedagóxica ou a posibilidade de limitar os resultados da procura. O involucramento do profesor e o bo deseño do corpus non son unha panacea, pero axudan a minimizar as dificultades que poida ter o usuario.

A recepción do DDL por parte do alumnado foi ben documentada nos últimos anos grazas a que o número dos estudos empíricos foi crecendo. No lado positivo, os estudantes reportaron que gozan ao traballar con linguaxe auténtica e gústalles que se poida ver o contexto do ítem en cuestión. Usaron os corpus como ferramenta de referencia para comprobar e confirmar o seu propio uso da lingua en cuestión. Doutra banda, ademais das críticas citadas anteriormente, algúns estudantes comentaron que lles parece moi tedioso examinar moitas liñas de concordancia e que lles resultaba difícil formular os termos de procura axeitados.

A recepción por parte dos profesores foi positiva pero cun certo grao de escepticismo. As preocupacións adoitan estar relacionadas co tempo necesario para crear exercicios de DDL e coa súa propia percepción como figuras de autoridade e coñecemento na aula. Moitos teñen un punto de vista pragmático que xira en torno á cuestión: *que proveito podo sacar disto, dadas todas as presións xa existentes?* Por tanto, non debería sorprender que os profesores poidan sentirse remisos con respecto á idea de traballar con corpus na aula.

Por iso, o concepto de *training* (adestramento) é moi importante. Non se pode esperar que un profesor ou un estudante saiba manexar un corpus sen recibir a formación necesaria. Iso non significa que esta formación teña que ser moi extensa, sobre todo se o corpus ten unha interface intuitiva. O corpus LITTERA conta cunha serie de exercicios de introdución para que o usuario se poida familiarizar cos textos do corpus e, coa súa composición, características e opcións de procura. Os exercicios forman parte dunha serie de vídeo-titoriais en Youtube creados polo autor. A ligazón aos titoriais atópase na páxina principal do corpus LITTERA. En menos de media hora, traballarían as habilidades necesarias para poder realizar as súas propias procuras no corpus.

O CORPUS LITTERA

Moi poucos estudos de DDL se centraron no tema da fonoloxía. Isto débese en parte a que, durante moitos anos, non había os recursos necesarios para crear un corpus con audio á escala necesaria para tal estudo. As limitacións tecnolóxicas das últimas décadas do século XX, xunto coa falta dos recursos humanos necesarios, non permitían que este campo florecese.

Para solucionar esta falta de corpus con audio (*speech corpora*) necesarios no campo de DDL, creamos o corpus LITTERA, un corpus paralelo literario audiotextual inglés-español deseñado para o estudo de diferentes aspectos da fonoloxía inglesa no marco do ensino do ESL (*English as a Second Language*). O corpus está composto por 25 textos literarios en lingua inglesa e as súas traducións ao castelán, así como segmentos de audio dos audiolibros correspondentes en inglés. Conta con case dous millóns de palabras (983.618 en inglés e 985.058 en español) e 63.508 unidades de tradución. Por cada unidade de tradución, hai tamén un segmento de audio que corresponde ao texto en inglés. Por tanto, hai 63.508 arquivos de audio. O seu nome débese ao dominio literario do corpus, xa que *littera* é a palabra latina de *letra*, e de aí, *literatura*.

Os textos foron seleccionados en gran parte partindo dos autores incluídos no Plan de Estudos do grao de Filoloxía Inglesa da Universidade de Santiago de Compostela. Na selección tamén influíu a dispoñibilidade do texto orixinal, a tradución e o audiolibro (en plataformas de uso aberto como Youtube ou LibriVox). Libros pertencentes aos xéneros de literatura infantil e xuvenil (como *The Hungry Hungry Caterpillar* e *Harry Potter and the Philosopher' s Stone*, respectivamente) foron engadidos para que houbese unha maior variedade de rexistros representada no corpus. O autor foi responsable da selección e aliñación dos textos, e a segmentación dos audiolibros. Tamén achegou ideas ás opcións dispoñibles no deseño, aínda que o traballo técnico foi realizado polo seu co-director de tese, Xavier Gómez Guinovart.

LITTERA forma parte de dúas coleccións de corpus, o corpus CLUVI e o corpus SensoGal, os dous aloxados no Seminario de Lingüística Informática da Universidade de Vigo. O corpus CLUVI facilita as consultas flexibles con expresións regulares, mentres que os corpus compilados no SensoGal están etiquetados semánticamente e permiten as procuras por forma (palabra ou lema) ou concepto (segundo os conceptos establecidos na rede semántica de WordNet e na súa adaptación en Galnet). A interface de SensoGal tamén permite realizar consultas especificando a categoría gramatical. As dúas coleccións ofrecen opcións para controlar o número de resultados, elixir a variedade oral do inglés (americano ou británico), ampliar o contexto, e, no caso de LITTERA, reproducir o audio dos resultados.

O corpus LITTERA está pensado para estudantes universitarios de fala hispana, aínda que calquera persoa que aprenda inglés pódese beneficiar del. Posto que os textos foron seleccionados a partir do programa de Filoloxía Inglesa, a maioría das obras son relevantes para os estudos universitarios. Ademais, é posible que a dispoñibilidade do texto orixinal coa tradución, xunto coa posibilidade de buscar dentro dun só texto se se desexa, anime ao estudante para ler textos que antes non tentaría ler pola súa dificultade.

AS APLICACIÓNS DIDÁCTICAS

As aplicacións didácticas desenvolvidas na tese céntranse en diversos aspectos suprasegmentales da fonoloxía inglesa, concretamente, en diferentes características da prosodia, tales como o *linking* (como as palabras se enlazan no discurso fluído), a *asimilación* (cando un fonema asimila as características dos fonemas ao redor), a *elisión* (a omisión dun fonema), a *palatización* (cando un fonema frontal, e.g. /t/ ou /d/, ocorre ante un fonema velar, e.g. /j/, e prodúcese un fonema africado como /tʃ/ ou /dʒ/), as *formas reducidas* (o feito de que moitas palabras gramaticais en inglés teñen unha forma de citación, i.e. de dicionario, e unha forma reducida para destacar as palabras que levan máis peso léxico) e a *unidade de entoación*. Aínda que o audio non provén de conversacións espontáneas, os aspectos da prosodia que se examinan aquí atópanse en todos os discursos do inglés falado. Por tanto, non é necesario que sexa linguaxe espontánea. Con todo, os audiolibros teñen a vantaxe de que as obras conteñen representacións de diálogo que reflicten a linguaxe conversacional, o cal, finalmente, permite unha análise dalgunhas das características do inglés conversacional tamén.

O primeiro aspecto da prosodia que se examina é o máis xeral, as unidades de entoación (*tone units*), que son unidades prosódicas con polo menos unha sílaba, que é a tónica da unidade, e que acaban cunha entoación de fronteira (a subida ou caída da entoación). As unidades de entoación están separadas frecuentemente por signos de puntuación, aínda que non en todos os casos. Poder recoñecer onde se sitúan as fronteiras entre unha unidade prosódica e outra é fundamental para logo entender por que certos fonemas realízanse de diferentes formas. As unidades de entoación pódense examinar con calquera procura, xa que cada unidade de tradución contén polo menos unha unidade de entoación.

O seguinte aspecto que se analiza é o concepto de *linking* (enlazamento), é dicir, como se conectan os fonemas a través das fronteiras entre palabras na fala. En concreto, examínanse as combinacións de C+V (cando unha palabra acaba en consoante e a seguinte empeza por unha vogal) e C+ C. Pódese atopar moitos exemplos no corpus para mostrar os diferentes fenómenos á hora de enlazar as palabras na fala fluída, tales como a elisión, a asimilación ou a resilabificación, que ocorre cando unha consoante na coda dunha sílaba convértese no ataque da sílaba seguinte. Isto analízase a través dos *phrasal verbs* no corpus, por exemplo, o fonema /k/ sendo resilabificado en *take off*, o cal pode causar a malinterpretación de *to cough*. Despois preséntase o concepto de *blending* para analizar os exemplos de C+ C cando as dúas consoantes son a mesma.

Unha vez establecidas as características básicas do enlazamento de palabras, examínase o que os datos mostran sobre as combinacións complexas de consoantes a través das fronteiras de palabras. O español non permite combinacións complexas de consoantes ao final dunha palabra, o cal xera moitas dificultades para os hispanofalantes ao atoparse con palabras como *asked* /æskt/. Máis dúbidas sobre a súa realización poden xurdir ao ter que combinar unha palabra xa complicada por si con outra palabra que empeza por unha consoante. Os datos do corpus mostran os casos nos que se pode reducir a combinación de consoantes a través da elisión de sons. Esta información é imprescindible non só para a percepción destes sons, senón tamén para facilitar a súa produción. É importante que os estudantes sexan conscientes de como poden simplificar as combinacións complexas.

A cuarta aplicación didáctica trata a asimilación a través das fronteiras de palabras. Os datos do corpus conteñen moitos exemplos deste fenómeno, que se pode atopar con facilidade grazas ao uso de expresións regulares. Un exemplo é a asimilación de /v/ ao final de *have* ante unha consoante xorda, como no caso de *have to*, converténdoa en /f/. Isto compárase cos casos nos que *have* é seguida por unha vogal, como no caso de *has a*, o cal produce unha consoante sonora /v/. As expresións regulares, por tanto, permiten buscar as dúas formas na mesma procura, o cal serve para comparalas e escoitar as diferenzas sen ter que realizar procuras separadas. Outro exemplo de asimilación é a palatización, que se pode apreciar nos datos do corpus, sobre todo en ítems de alta frecuencia como *what you* e *but you*, xunto con outros exemplos nos que un narrador elixe facela con algúns personaxes e non con outros, dando lugar a unha variación estilística motivada no discurso.

Coa análise dos fenómenos fundamentais da prosodia—asimilación e elisión de sons—, xa existe unha base para examinar a realización dun morfema específico, o do pasado simple, *-ed*, como en *work**ed***. Aínda que a pronuncia deste morfema, /t, d, ɪd/, adóitase ensinar segundo a sonorización do fonema anterior (nos casos de /t/ e /d/) ou o punto de articulación (no caso de /ɪd/), os datos do corpus demostran que hai outros factores que contribúen á súa realización, como os fonemas posteriores e a presenza dunha fronteira dunha unidade de entoación. Por iso, é frecuente que os estudantes non perciban a realización deste morfema debido a estes outros factores prosódicos, como a elisión e a asimilación. Isto pode causar unha transformación inesperada no morfema que non permite a súa percepción se un só espera escoitar os fonemas /t, d, ɪd/ da regra orixinal. Por tanto, nesta sección propóñense unha serie de preguntas baseadas no DDL para encamiñar o estudante cara ao descubrimento dos datos reais que poidan informar a súa propia comprensión e produción deste morfema.

A continuación, explórase a redución vocálica nos datos do corpus. Un dos exemplos principais é o do verbo auxiliar *can*. É común que os estudantes se queixen de que moitas palabras parecen ‘desaparecer’ na fala fluída en inglés. Isto débese a que moitas palabras funcionais/gramaticais teñen polo menos dúas formas, a enteira (i.e. de dicionario) e a reducida (influenciada polas características da prosodia). Os datos do corpus mostran as dúas formas e como *can* se pode reducir e por que *can’t*, en cambio, nunca se reduce. Tamén se pode apreciar certos casos nos que *can* non se reduce para mostrar énfase ou contraste. Esta distinción entre as dúas formas é fundamental, xa que ao usar a forma enteira de *can* coa vogal aberta, é posible crear confusión para o interlocutor e que interprete *can* como *can’t*, sobre todo porque a elisión do fonema /t/ en posición final é algo frecuente (e que se pode apreciar nos datos). Por tanto, os estudantes deberían ser conscientes das dúas formas para poder comprender un enunciado con *can* adecuadamente, e saber utilizalo para non crear confusión con *can’t*. Tamén se examinan as formas reducidas do verbo auxiliar *have*. Tamén se presenta aquí o concepto de clitización (por exemplo, achegar *have* a outro verbo auxiliar, como *should* para producir unha forma clitizada de *have*: *should’ve* ou *shoulda*).

Dado que o corpus está composto por textos literarios, a seguinte aplicación didáctica trata sobre a representación da linguaxe falada no diálogo dos textos. Aínda que LITTERA non contén audio de conversacións espontáneas, o diálogo escrito serve como un substituto diso. A pesar de que os autores non adoitan ser lingüistas profesionais, en moitos casos tentan captar a forma ‘auténtica’ ou ‘realista’ na que falan as persoas. Os autores si son falantes do idioma e queren reflectir eses matices da fala conversacional que poidan servir para informar o lector sobre o personaxe en cuestión. Por tanto, moitas características do inglés falado e conversacional aparecen nos datos e sinalan información importante sobre os efectos da prosodia. Isto inclúe representacións escritas de características conversacionais tales como a clitización (o cal inclúe contraccións) e a redución de palabras (por redución vocálica ou elisión). Algúns exemplos destes fenómenos representados no diálogo escrito inclúen as contraccións dos verbos modais (e.g. *she's gone*, *there'll*, *what'll*, *it'd*), as formas clitizadas de *have* (*kinda*, *sorta*) e *to* (*gotta*, *gonna*, *wanna*), a elisión de sons ao comezo da palabra (*'em*, *'ud*, *'im*, *'as*, *'is*, *e 'em*), e a redución de *for* a unha vogal media en *fer*. Todos estes exemplos reflicten tendencias reais do inglés falado. Esta información é extremadamente útil para a comprensión e produción natural do idioma.

Por último, analízase un aspecto característico do inglés norteamericano: o *flap* (tamén coñecido como o *tap*, ou 'toque'). O *flap* ocorre cando unha consoante coronal /t/ ou /d/ aparece nunha posición intervocálica. Ocorre principalmente no inglés norteamericano, aínda que se pode atopar noutras variedades de forma menos extensa. O *flap* pódese atopar dentro da palabra ou entre dúas palabras, sendo o segundo caso o máis complicado para os estudantes, xa que é menos predicible. É máis, o *flap* pode crear homófonos que noutras variedades non o son, como é o caso de *seeded* e *seated*. A través da opción de buscar no corpus por variedade falada (*Spoken English Variety*), pódese apreciar estas diferenzas na pronuncia. Aínda é máis complicado cando o *flap* ocorre entre varias palabras seguidas, e.g. *le**t** i**t** out* e *pu**t** i**t** out*. Esta sección demostra como o profesor pode explorar este fenómeno no corpus cos estudantes, sobre todo cos ítems de alta frecuencia como *that* + V (e.g. *that everything*) ou *get* + V (e.g. *get up*), para que sexan conscientes deste fenómeno en certas variedades do inglés.

COMENTARIOS FINAIS

O presente traballo describe a creación e o deseño do corpus LITTERA e presenta algunhas aplicacións pedagóxicas para o estudo da fonoloxía inglesa desde a perspectiva teórica de DDL no ámbito do ensino do ESL. O corpus LITTERA é un corpus paralelo literario audiotextual inglés-español pensado para estudantes universitarios hispanofalantes de filoloxía inglesa, xa que os textos literarios, en gran parte, foron seleccionados en base ao plan de estudos para ese grao na Universidade de Santiago de Compostela.

O DDL é unha aproximación á aprendizaxe de idiomas na que os datos do corpus serven como *input* e substitúen á forma tradicional de recibir a información do docente de xeito pasivo. Desta forma, os estudantes asumen un papel activo na súa aprendizaxe, o cal conleva unha serie de vantaxes, tanto a nivel lingüístico como persoal. Posto que a maioría dos traballos de DDL céntranse nos corpus textuais, o autor ten a esperanza de que esta tese doutoral axude a achegar os estudos de DDL aos corpus con audio (*speech corpora*).

Este traballo mostra como un corpus con segmentos de audio pode servir como ferramenta para a aprendizaxe da fonoloxía inglesa nun contexto de ESL. Analízanse diferentes aspectos suprasegmentais do inglés que non se adoitan estudar demasiado coas metodoloxías tradicionais. Desta maneira, os estudantes, coa axuda do docente, poden facer as súas propias observacións sobre os aspectos prosódicos da fala fluída, como a redución vocálica, a elisión de sons, a asimilación, a clitización e a palatización, entre outras tendencias. Desas observacións pódense chegar a formar regras que axudarán na comprensión oral e na produción.

O autor espera que esta investigación dea lugar a estudos empíricos que permitan comprobar a eficacia deste corpus, ou calquera outro con audio, na aprendizaxe da fonoloxía na aula. As aplicacións didácticas elaboradas aquí teñen como obxectivo demostrar o que din os datos do corpus e por que paga a pena que os estudantes traballen con eles. A pesar de que o campo de estudo do DDL creceu moito nos últimos anos, segue habendo moito territorio sen explorar, do cal esta tese doutoral tentou alumar unha parte.

INTRODUCTION

We now live in an era in which, thanks to the widespread and ever-expanding nature of the Internet and the *smart* devices we use to access it, a vast amount of language learning is done in the digital realm through apps, streaming services and online learning platforms. As each generation seems to become more tech-savvy than the last, more and more students appear to be turning to digital resources over textbooks and traditional methodologies. As a result, recent decades have seen a growing interest in shifting linguistic technologies, particularly online corpora and concordancers, from the realm of expert researcher to that of the non-expert language learner. The evidence for such a shift can be found in the rise of Data-Driven Learning (DDL), an approach to language learning in which the student interacts directly with corpus data rather than indirectly through more traditional mediums such as textbooks and dictionaries.

Most DDL research has been focused on textual corpora, which means that aspects of language learning such as speech production and oral comprehension have been largely absent from the literature. Due to this gap in the research, the author and his doctoral supervisor have created the LITTERA corpus, an audio-textual English-Spanish parallel literary corpus. One of the main goals in LITTERA's inception has been to establish a freely available and user-friendly speech corpus that can be readily accessed and explored via the web interface, rendering it an optimal pedagogical resource in the study of English phonology, among other things.

Therefore, the aim of this dissertation is two-fold: firstly, to introduce and describe the LITTERA corpus by detailing its conception, creation, composition and features; and secondly, to examine in what ways the corpus can be conducive to the learning of certain aspects of English phonology, as it is the author's contention that this area of study has been underrepresented in the DDL literature. It is hoped that the pedagogical applications described here lay the groundwork for future empirical research to be carried out in order to assess LITTERA's effectiveness as a DDL resource — or that of any other similar corpora —and how it will be received and utilized by students of English.

In chapter 1, the Data-Driven Learning approach is presented along with an overview of the current state of research. Chapter 2 describes the LITTERA corpus at length, including its conception, compilation, current features and search capabilities. Chapter 3 presents a series of potential pedagogical applications of the corpus in language learning, particularly suprasegmental features of English phonology, before final remarks are made regarding the research presented here as well as future research as it pertains to LITTERA and DDL as a whole.

1. DATA-DRIVEN LEARNING: AN OVERVIEW

Since its inception in the 1980s, a fair amount of research has been carried out in the field of DDL. Tim Johns, one of its earliest pioneers, explained the methodology behind DDL as follows:

> "What distinguishes the DDL approach is the attempt to cut out the middleman as far as possible and to give the learner direct access to the data, the underlying assumption being that effective language learning is a form of linguistic research, and that the concordance printout offers a unique way of stimulating inductive learning strategies – in particular the strategies of perceiving similarities and differences and of hypothesis formation and testing" (Johns, 1991b)

A more succinct version of this can be found in his now widely cited "every learner a Sherlock Holmes" (Johns, 1997). The once simple idea of students as "language detectives" has blossomed into a significant body of research. DDL continues to expand as we shift from (but not abandon) the concordance printouts referenced by Johns to more student- and user-friendly electronic corpora.

DDL has been credited with such direct effects on language learning as improvements in vocabulary (Römer, 2008), recognition of lexico-grammatical patterns (Sripicharn, 2010; Smart, 2014) and regular exposure to high-frequency items. Furthermore, rule-teaching in traditional pedagogical contexts often falls short in providing a complete picture of how the language is actually used and tends to be very general, vague, or abstract (Boulton, 2009). By analyzing true instances of language use, students are able to examine the variety of contexts in which a word may occur, leading to a more well-rounded understanding of its different uses and meanings. DDL has also been credited with increased learner autonomy and an improvement of overall language awareness (Talai & Fotovatnia, 2012) and noticing (Boulton, 2011). Students can bring their own research questions to the task and use the corpus to further refine their understanding of a given aspect of the language. By interacting directly with the corpus data, students take an active role in their own language learning (Chambers, 2010), rather than the traditional model of passively receiving information from an instructor. Such direct participation allows for the conceptualization of learning as a process of discovery (Bernardini, 2004) in which students identify patterns in the language for themselves without direct teacher intervention. While one might suspect that only advanced language learners would benefit from the complex and often fuzzy nature of authentic language use, researchers have shown that DDL can be successful throughout all levels of language learning (Boulton, 2017). Finally, Boulton & Cobb (2017) provide a nice summary of the many benefits of DDL:

> "The DDL approach is geared to making sense of language input but has several potential advantages that other input approaches do not. Core among these is that input assembly replaces input simplification, thus maintaining authenticity of language. Another advantage lies in identifying which forms and meanings in a language (whether words, structures, pragmatic patterns, etc.) are most frequent and thus probably most worth knowing. DDL consists of the consultation of language data by learners themselves and thus incorporates the notions of learner autonomy, induction, exemplar-based learning, and constructivism, in the sense of letting learners discover linguistic patterns for themselves (with varying degrees of guidance) rather than being spoon-fed predigested rules."

Despite boasting what is now a substantial body of research that continues to expand, DDL is still a far cry from being truly commonplace in language classrooms. Lee (2011) lays out some of the reasons for this, which include the instructor's overall familiarity and level of comfort with corpora, the time-consuming realities of preparing relevant pedagogical material, tailoring the data to students' needs, and, of course, the curricular requirements that must be met by teachers. Proposed solutions to these issues include pedagogically motivated corpora aimed to meet learners' needs (Braun, 2005), textbook corpora and graded-reader corpora (Lee, 2011) and local learner corpora (Mukherjee & Rohrbach, 2006). Lee sums it up adequately stating that, "the appropriate and effective use of corpora in the classroom is partly a technical issue, but primarily a pedagogical one". In other words, even when technological barriers are a non-issue, there is still the underlying matter of teachers and students understanding the nature of corpora, the variety of corpora that exist, which ones are more readily available and easily accessible, which are better suited for language learning and their specific needs, and, perhaps most importantly, what knowledge can be gained with corpora.

1.1 WHAT IS A CORPUS?

Before we continue our analysis of DDL, we must first take a step back and discuss its pedagogical centerpiece, the corpus. While it is easy to think of a corpus as simply a collection of examples of language use, McEnery et al. (2006:5) point out that a corpus has more precise characteristics which distinguish it from just any random collection of texts, stating that "there is an increasing consensus that a corpus is a collection of (1) *machine-readable* (2) *authentic* texts (including transcripts of spoken data) which is (3) *sampled* to be (4) *representative* of a particular language or language variety". Definitions provided by other researchers overlap with McEnery et al.'s, as in Sinclair (2005), who describes a corpus as "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research", or Leech (1992:116) who describes a corpus as text collections that are "generally assembled with particular purposes in mind, and are often assembled to be (informally speaking) *representative* of some language or text type". In other words, a corpus is a collection of texts that has been compiled with a specific objective, and to fulfill said objective, it is necessary to make considerations for such aspects as the composition of the corpus and what type of language the texts are indicative of, i.e. their representativeness (see section 2.6 below for more on representativeness). In other words, the texts that make up a corpus for linguistic purposes are not there by accident. Furthermore, in modern-day Corpus Linguistics, these texts are accessible electronically, which has allowed the field to grow and branch off into other areas, such as language pedagogy.

Corpora have long had a role in language pedagogy, albeit an indirect one, by providing real language data for pedagogical resources such as dictionaries and reference grammars (McEnery & Xiao, 2011), a practice that continues to this day. DDL, then, provides an approach that brings students and teachers in direct contact with corpora.

While print-outs and other physical forms of corpus data are still used in the field, the shift to electronic corpora has been significant since DDL's early days and allows for a more comprehensive interaction with the data than could be achieved on paper. This is thanks to features such as corpus annotation (part of speech tagging, lemmatization, semantic tagging), multimodal corpora (the addition of other media, such as audio and video), concordancing software, the ability to carry out more sophisticated search queries (via the use of regular expressions) and overall improvements in user-friendliness (well designed web-interfaces). All of these features have given rise to multi-faceted corpora that are often freely available online at students' fingertips.

1.1.1 Different Types of Corpora

As the field of Corpus Linguistics continues to grow, so do the different types of corpora created for both research purposes and language pedagogy.

Text is by far the most ubiquitous data type, which comes as no surprise given that it is the most readily accessible and easiest to compile. Whether it comes from written or spoken sources depends on the aims of the compilers of a given corpus. The text may be arranged in a variety ways and include different languages. For monolingual corpora, the text structure is generally straightforward, whereas with multilingual corpora, the texts may be translations or simply share a common domain (e.g. biblical texts, travel brochures, etc). In the case of translations, it is common for the texts to be presented "parallel" to each other, often aligned at the sentence or paragraph level. Parallel texts may be unidirectional or bidirectional, depending on the source text for the translation(s). When the texts share the same domain rather than existing as translations, this is commonly referred to as a comparable corpus. Domain simply refers to the topic or theme of the corpus. The domain of a given corpus may be specific, such as a corpus made up of editorial articles on the invasion of Iraq from 2003 to 2010, or broader, such as a corpus of 20th century Irish fiction.

The text may also be annotated in some way to provide additional information for more refined search capabilities, such as the signaling of errors in a learner corpus or part of speech tagging. Annotation depends on the purpose for which the corpus was designed and is determined by the research or learning objectives. Thus, a corpus to be used for translation research will most likely have different annotations than, say, a learner corpus.

Beyond text, other types of data are becoming more commonplace, such as audio and video, thanks to changes in technology that have made it easier and less costly to add new forms of media to a corpus and make them accessible to users. As a result, there has been a growing number of speech corpora used for both research and pedagogy in recent decades. Unlike a spoken corpus, which typically limits its scope to the written transcriptions of spoken communication, a speech corpus provides actual audio recordings, also regularly accompanied by a transcription, whether orthographic, phonetic, or both. Such multimodal corpora may be used in a number of different areas of research, from pragmatics, to discourse analysis, to phonetics and to language learning.

Finally, representativeness is a frequently cited principle in corpus design methodology (Guan, 2013) and, of course, is contingent upon the domain. Does the content paint an accurate picture of the domain in question? Are the language samples truly representative of, say, 19th century Irish non-fiction, or Modern Academic English? As Mukherjee (2006) explains, “representativeness is a key issue in corpus design because it captures the attempt to compile a database that provides a statistically viable sample of language use in general”. However, as Braun (2006) points out, “the ways in which corpora are used in a pedagogical context differ from linguistic research contexts” and that for pedagogical corpora, “homogeneity and topical relevance are more important than representativeness in the traditional sense”. This distinction would lead her to call for more “needs-driven corpora” (Braun, 2007), which will not follow the same corpus methodologies for compilation as research corpora. Therefore, the aims of the corpus must be decided before questions of representativeness can be addressed. Section 2.6 below will provide more information on representativeness as it pertains to the LITTERA corpus.

1.2 DDL: STATE OF THE ART

The body of literature surrounding DDL continues to expand, although with the vast majority of research being carried out within English learning contexts (Vyatkina, 2016a). Particularly in its early days, DDL has received much criticism for an apparent lack of empirical research to support the optimistic claims made by its proponents. Nowadays, this is no longer the case, with

Boulton (2017) even describing DDL's current research situation as "flourishing" in his meta-analysis of DDL studies.

While many descriptions abound, DDL as it is understood today may best be summed up as a "student-centered method in which natural instances of language produced by native speakers are gathered and presented to the learners for the purpose of improving language proficiency" (Talai & Fotovatnia, 2012). This is accomplished through the principle of *noticing* (Schmidt, 1990), which argues that in order for an item to be learned, learners must be consciously aware of said item or construction. In his own words, Schmidt (1995) simply describes *noticing* as the "conscious registration of the occurrence of some event". By interacting with corpus data, usually through concordance lines, learners become aware of certain salient features that function as a form of language input. It is this input awareness that Schmidt (2001) argues is essential to the language-learning process, for it gives way to understanding. Schmidt (1995) goes on to explain that, "noticing refers to surface level phenomena and item learning, while understanding refers to deeper level [sic] of abstraction related to (semantic, syntactic, or communicative) meaning".

What sets DDL apart from more traditional language learning approaches is that students interact with "authentic" language data, or Talai & Fotovatnia's "natural instances of language produced by native speakers", rather than simple, neat and contrived examples often used to demonstrate certain grammatical structures. Johns (1991a) described the data in corpora as "the facts of linguistic 'performance'". This falls in line with the view of certain researchers such as Sinclair (1991) and Kennedy (1992) who dismiss the idea of invented examples providing language learners with an accurate depiction of real language use, arguing that such examples distort and mislead due to their reductive nature.

While the idea of authenticity seems rather uncontroversial, it has nevertheless attracted a certain amount of debate among DDL researchers, leading some to avoid the term "authentic" altogether. While many researchers have viewed authenticity at face value, falling in line with Talai & Fotovatnia's description, some have taken issue with the idea of authenticity as an inherent part of a text. Widdowson (2000) was the first to distinguish between how a text is created and how it is received, arguing that by removing a text from context in which it was conceived, what is authentic for the creator of a text, even if it is a genuine instance of language use, is not necessarily authentic for the receiver, i.e. the language learner, and that it is necessary for students to be able to "authenticate" a text for themselves. In other words, because a text in a corpus has been stripped of its original context, a learner must re-contextualize it and establish a new relationship with it (Boulton, 2009) so that it is authentic for him or her. Braun (2005) suggests that this can be achieved from different angles by taking into account factors such as content, design (she specifically names "size, data format and annotation"), as well as how the data is analyzed.

DDL has also received its share of criticism. The broadest criticisms are usually related to technology, specifically in terms of access and personal comfort. If much of DDL is now done online, this presents a problem for those classrooms that do not have quick and easy access to computers or a wireless internet connection. This can be remedied, to some extent, through the use of printouts. Studies have found no notable differences in the learning outcomes between those whose DDL tasks were computer-based and those whose tasks were paper-based (Boulton, 2008; Boulton, 2012; Vyatkina, 2016a). Students may also have an aversion to working on a computer and prefer physical printouts (Boulton & Cobb, 2017). This becomes more problematic if teachers are the ones who do not feel at ease with computer-based work since they are responsible for preparing the activities. This is an undeniable problem for our purposes here, as LITTERA only exists as an online platform and relies on the digital audio files taken from the audiobooks, rendering this aspect of the corpus incompatible with paper-based methods. However, as younger generations now grow up practically immersed in various forms of technology, it is likely that this aversion to computers and online learning will become a less significant issue in the years to come.

Another valid criticism is the lack of corpus knowledge on the part of teachers. Many of the surveys that have been carried out to learn about teachers' familiarity with corpora tend to show that only small percentages of teachers receive exposure to corpora and corpus methodology at some point throughout their undergraduate and certification programs (Mukherjee, 2004; Breyer, 2009; Frankenberg-Garcia, 2012; Tyne, 2012). Language teachers are seldom experts in corpus linguistics. Thus, some training is normally required in order to able to create effective DDL tasks for students (see section 1.2.4 below for more information on training).

Even if such training is possible, DDL may not be practical for those teaching an already bloated curriculum as many have criticized it as being too time-consuming (Yoon, 2011; Boulton

& Cobb, 2017). Teachers are frequently required to teach a large amount of content in a small space of time, especially at the secondary and university level. Therefore, designing and executing DDL activities may not be deemed realistic in certain circumstances. Teachers may need to locate and extract the concordance examples they want to use in order to tailor the data to the needs of the activity. Some teachers might simply view this as too much preparatory work just to examine a handful of instances of a word or expression. Added to this is the fact that some degree of teacher supervision is recommended to prevent learners from drawing inaccurate conclusions from the corpus data.

Finally, more specific criticisms of DDL usual revolve around corpora and concordancers. Although many corpora are accessible online through web interfaces with their own built-in concordancing tools, there are still many cases in which concordancing software, such as AntConc[1] (Anthony, 2015) and Lancsbox[2] (Brezina et al., 2018), will be necessary (e.g. when using local corpora). Such corpus software, and even some online interfaces, present the concordance data in incomplete, chopped-off concordance lines that many consider too difficult to make sense of. This comes back to the issue of authentication, for as Braun (2006) notes:

> "…the question is how easy or difficult it is for learners to authenticate a text which they 'access' through a concordance line — without knowing to which part of the text this takes them and without pedagogically appropriate information about the text and the communicative situation in which it was produced."

Another issue that has been brought up is that the native-speaker language contained in the corpus may be too difficult for learners (Boulton & Cobb, 2017), or the data may be too overwhelming for learners to make sense of. If hundreds or thousands of results are returned for a search item, where is the learner to start? Yoon (2011) has pointed out that the exact opposite could occur and there are not enough examples to get a clear idea of how a word or expression is to be used. These issues can be addressed, at least in part, through proper corpus design. Through

[1] https://www.laurenceanthony.net/software
[2] http://corpora.lanc.ac.uk/lancsbox
[3] The asterisk is functioning as a regular expression meaning any number of characters in that space. See section 3.1.1.3 below for
[2] http://corpora.lanc.ac.uk/lancsbox

such techniques as pedagogical annotation or the ability to limit search results, the problems may not completely go away, but their impact on learning may be significantly reduced.

1.2.1 DDL and Language Pedagogy

DDL has been shown to be well grounded in the principles of Second Language Acquisition (SLA) and learning theory (Gilmore 2007; Boulton & Tyne, 2013; Vyatkina, 2016a & 2016b; Boulton & Cobb, 2017). As the idea of *noticing* has already been introduced in section 1.2, the following will examine the other ways in which DDL corresponds to these well-established pedagogical principles.

Advocates of DDL have cited numerous advantages over traditional language teaching methods, many of which were mentioned in the introduction to this chapter. One reported benefit recurrent in the literature is that of *inductive learning*, which stems from the aforementioned pedagogical principle of noticing and is directly linked to the idea of learner autonomy. In fact, Vyatkina (2016b) points out that the term "guided induction (or guided discovery), originates in general language acquisition theory". Inductive learning is when students make inferences based on what they find in the data, rather than beginning with an explicit explanation of a rule. This leads students to take on greater responsibility for their own learning, and therefore increased autonomy, as it puts them in a position where they must find patterns and work out their own rules rather than being passively fed the information by teachers. This notion is sometimes referred to as bottom-up learning (Mishan, 2004). What makes this type of learning effective is that learners are more likely to remember the rules they themselves have formulated and discovered on their own (Jones, 1997). McEenery & Xiao (2011) explain the "three stages of inductive reasoning with corpora in the DDL approach" originally proposed by Johns (1991a); these are "observation (of concordanced evidence), classification (of salient features) and generalization (of rules)". That is, by observing the data, the student can begin to recognize patterns that eventually lead to rule formation.

Boulton & Tyne (2013) give a summary of the many claims made in the research that stem from this type of student-centered learning:

> "DDL is alleged to enhance cognitive and metacognitive skills, increase sensitivity to authentic language use, provide an interactive approach to constructivist discovery learning, foster motivation especially through individualization, promote reusable and transferable skills, favour autonomy for life-long learning, and correspond largely to current theories of second language acquisition"

Constructivist theory here refers to "learning as a constructive process in which the learning is building an internal illustration of knowledge, a personal interpretation of experience" (Guan, 2013). Boulton & Tyne (ibid), when comparing DDL to more "direct" input methods, go on to state that "a constructivist approach allowed greater immediacy, personalization, involvement, and any number of incidental benefits which are difficult to assess in a traditional research paradigm".

All this is not to say that deductive, top-down learning is to be left by the wayside. Even Johns (1991a) recommended that DDL should work alongside "older and more familiar methods". This is because DDL is also compatible with more traditional learning styles based on direct, rule-driven instruction from a teacher or textbook, i.e. *deductive learning*. For example, a corpus may be used to falsify a certain grammatical rule being taught in the classroom. In doing so, students are following the steps of basic scientific inquiry; that is, hypothesis formation, followed by experiment and observation. In the same vein, learners may also use a corpus to check or correct possible errors in their own language use (Sripicharn, 2010) and to prove or disprove their own assumptions about the target language, all of which requires significant autonomy on the part of the student. This can also serve to correct errors that students repeatedly make, so-called fossilized errors (Nesselhauf, 2004).

Learner autonomy may also be increased through what Bernardini (2004) calls "serendipitous learning". This refers to learning done beyond any organized DDL activity. Here, students have been found to discover unexpected features of the language not necessarily related to the task at hand through simple corpus browsing, letting their own curiosity guide them. This kind of unsupervised learning promotes independence and may increase motivation as students are consulting the corpus for issues that are immediately relevant to their needs and based on questions that arose organically. There is a danger in this, however, as students may arrive at the wrong conclusions when forming rules from the data without any supervision. They must learn to be cautious when working completely independently as there is always the possibility of a misinterpretation of the data (see section 1.2.4 below on training for further discussion on how to prepare learners to deal with these issues).

Pattern recognition, fundamental to DDL, is another important aspect of learning theory (Boulton & Cobb, 2017). Pattern recognition is what leads Talai & Fotovatnia (2012) to point out that the DDL approach is similar to learning a language naturally since the learner must decipher patterns from natural examples from real speakers of the language without the luxury of being taught exactly *why* certain structures are the way they are beforehand. Chambers (2010) claims that, as a result, "[g]iving learners access to multiple examples of common patterns could help overcome what Debrock *et al*. (1999:46) call '*le manqué de naturel* [the lack of naturalness]' in learner language". In addition, Granger and Gilquin (2010) claim that DDL can lead to a "heightened awareness of language patterns".

DDL, in short, allows students to discover patterns in the language through corpus data and formulate their own rules, which is what all language learners already do at conscious and unconscious levels whenever they interact in the L2, especially with native speakers. This methodology coincides with many principles in SLA and general learning theory, specifically inductive (or discovery) learning via pattern recognition, which promotes motivation and learner involvement and leads to greater learner autonomy. Nevertheless, it is important to keep in mind that what works for one student may not work for another. Success with DDL is not an immediate guarantee and one must take into account a range of factors, such as how the activities are designed, the amount of guidance needed/given from teachers, corpus design, and, of course, the influence of personal affect on an individual's learning. Researchers regularly warn of the ease with which one may assume DDL to be a kind of language learning panacea, and while empirical results have been positive (see Boulton, 2017), one cannot overgeneralize the approach and assume it will work for all learners in all pedagogical contexts.

1.2.2 The Role of Teachers

The bottom-up nature of DDL requires us to re-examine the roles of students and teachers in a language classroom. As discussed in the last section, DDL allows students to take on more responsibility for their learning and rely less on direct, explicit instruction from the teacher. This then begs the question: if the student is becoming more autonomous in his or her learning, what does that mean for teachers? Any major change in the age-old dynamic, especially one in which students are finding the answers in technology rather than the wise, all-knowing instructor may lead some teachers to worry that they will slowly be replaced by machines at worst, or that this undermines their authority at best. McCarthy (2008) describes teachers as often "frightened" by the possibility of introducing corpora into the classroom because that may imply that the teachers should also be computational linguists and/or native speakers to be able to make the most of corpora. McCarthy goes on to note that for too long teachers have been seen as "consumers" of corpus-based materials, e.g. dictionaries and textbooks, and asks how we can "turn that 'consumer' into a more active participant in the corpus revolution".

Johns (1991b) envisioned the role of the teacher shifting from that of "expert" to that of "research organiser". That is, the teacher is no longer viewed as a bearer of special knowledge that is then passed down to the student via explicit instruction, but rather as a facilitator who guides students through the process, from formulating queries to the analysis and interpretation of the results (Chambers, 2010). As Lee (2011) puts it, "it is crucial for teachers to make corpus data 'palatable' for the learner by 'preselecting' and 'digesting' raw corpus data in order to make them pedagogically more relevant". Thus, the teacher's role shifts from directly imparting knowledge to that of providing students with adequate learning materials so that they can then discover patterns in the language more autonomously.

Granger & Gilquin (2010) describe DDL activities as existing on a "cline ranging from teacher-led to learner-led. On the learner-led extreme, we find Bernardini's "serendipitous" corpus browsing, in which students take on a high level of autonomy through the aforementioned discovery learning. This has also been referred to as the "hands-off" approach. At the other extreme we find highly controlled activities using corpus data in which students are working to answer specific questions proposed by the teacher. This has been described as a "hands-on" approach. Vyatkina (2016b) found that both hands-on and hands-off approaches were successful in learning L2 English and L2 German collocations.

The degree to which the activities are more hands-on or more hands-off depends first and foremost on the learning context. Granger & Gilquin (ibid) recommend the hands-on approach for beginners as they are still building the foundations of the language while more advanced learners will benefit from hands-off learning as they are merely locating and filling in gaps in their knowledge. Nevertheless, Vyatkina (2016a) found success by applying the hands-off approach with low-intermediate learners of German.

Teachers can also benefit directly from corpora because corpora can inform them as to which items are high-frequency and worth spending class-time on, much in the same way as how corpora can inform dictionaries and textbooks. Furthermore, many language instructors are teaching their L2 and may have certain doubts that only native-speaker consultation can resolve. Corpora are ideal for these circumstances (Römer, 2011) and can even provide much more information than a native speaker would be able to as corpora are usually full of instances of language use by a *variety* of native speakers, not just one.

Chujo & Ohigian (2008), who advocate for combining both inductive and deductive approaches, provide some examples of possible tasks teachers can use in DDL activities for learning noun and verb phrases at the beginner level. To explore “different verb forms and derivations”, they formulate the prompt as: “Search ***develop**** and list both different verb forms and derivations[3]” (bold and italics from original). The answer is then given as “*develop*, *develops*, *developed*, *developing*, and *development*”. To examine “the basic structure of VPs”, the task is to “[s]earch ***enjoyed*** and find which verb form frequently follows it. *(Answer: to-infinitives and gerunds.)*”. While these types of activities undoubtedly require a certain amount of preparation on the part of the teacher as he or she would have to make sure the answers are clear enough from the data, the resulting activity is very straightforward and will allow students to become acclimated to interacting with corpora.

Ultimately, it is the teacher who is responsible for the role of bringing corpora into the classroom, a role which Breyer (2009) argues has not received much attention at all. She notes specific factors underlying the decision of whether or not to introduce corpora into lessons, such as “motivation, availability of materials and the possession of adequate skills to teach with corpora”. While it is hard to change another’s motivation, by creating more available resources and providing the basic training in order to generate a certain level of comfort with corpora, it is

[3] The asterisk is functioning as a regular expression meaning any number of characters in that space. See section 3.1.1.3 below for more details on regular expressions.

likely that these pragmatic changes will also bring about the impetus needed for active engagement with corpora in the classroom. As will be seen in the following chapters, attempts have been made to reduce these obstacles as much as possible in the case of LITTERA through both corpus design and the availability of introductory tutorials for corpus work.

All this considered, it is clear that fears of replacing teachers with machines are unfounded as the teacher's role is no less important now than before, and that it is only shifting in terms of preparation and responsibilities. As is the case with many innovations in education, rather than worrying about becoming obsolete, teachers must adapt to new methodologies and find ways to make the tools at hand most beneficial for their students.

1.2.3 Student and Teacher Perceptions

Perceptions of DDL have been fairly well documented in recent years as empirical studies have become more widespread. In an overview of the research, Yoon (2011) describes many of the positive and negative reactions students have had to working with corpora. On the positive side, students seemed to enjoy working with "authentic language" and liked that the corpus provides the "contexts where words and structures are used". Students used corpora "as a quick and easy reference for checking and confirming". They also claimed to feel a sense of "greater autonomy in learning" and "confidence in L2 writing".

Conversely, negative reactions, which echo those already mentioned in section 1.2 above, describe DDL as too "time consuming to sort through concordance examples and identify relevant ones" and note cases where the results provided "too few or no incidences of the search item". Other negative reactions cited by Yoon include frustration from "not [being able] to understand all concordance examples" and that it was "hard to formulate proper search terms".

Boulton & Tyne's (2013) own overview of the research confirms many of the perceptions reported by Yoon, adding, on the positive side, that many learners showed great enthusiasm and reported their intention to continue using corpora in the future in their own independent work. On the other hand, in his timeline analysis of DDL research, Boulton (2017) suggests that maybe more research should be carried out on "DDL and learning styles, strategies and motivations" and also noted, in agreement with Yoon, the difficulty some students have with formulating searches adequately.

As for teachers' perceptions, Mukherjee (2004) surveyed 248 German teachers of English regarding their knowledge and perceptions of corpus linguistics and its applications to language pedagogy. This was done before and after what he calls *test workshops*, meant to introduce secondary level teachers to the principles of corpus linguistics and what they can mean for language learning. Going into the workshops, around 80% of teachers surveyed had never encountered corpus linguistics before. At the end of the workshop, after being introduced to corpus techniques and DDL activities, the vast majority (over 95%) viewed corpora as advantageous for English language learning, but with the caveat that most felt they were not suitable for use by the students, only the teachers. Mukherjee suggests that:

> "this sheds light on an important clash between applied corpus-linguistic research and the average teacher's point of view; while in applied corpus linguistics, there is an increasing tendency to focus on corpus-based activities carried out by increasingly autonomous learners…most teachers think that corpus data are particularly useful for themselves".

Finally, when asked which activities they would consider putting into practice in their own classrooms, teachers "exclusively focused on teacher-centered activities and showed that learner-center activities have no place in their classrooms".

Breyer (2009) reports on the perceptions of a group of student teachers, also from Germany, most of whom had no prior knowledge regarding corpora or the field of corpus linguistics. The student teachers were introduced to corpora and asked to carry out their own data-driven tasks on the English topic of "some vs any". After exploring "some" and "any" through concordance lines, "[t]he student teachers recognised the value of the corpus and concordance as tools for the learner to explore the complexities of language and also to lend credibility by allowing the learner to explore authentic text and discover language use at their own pace". The tasks "led not only to an increase of language awareness but teaching awareness as well". However, these positive reactions were not without their concerns. When asked to prepare DDL exercises of their own for their students, they found that "finding a suitable corpus proved to be the most difficult part of the assignment". This is also complicated by the inherent *messiness* of natural language use. Breyer notes that they “were looking for examples to support their desired learning target rather than language samples that would reflect language use as it naturally occurs”. Furthermore, “the student teachers viewed the unpredictable nature of concordancing exercises as a source of concern rather than an asset”. In short, while the student teachers saw value in direct corpus exploration, it appeared that they favored resources that provided more black and white solutions to the language topics they cover in class. Corpora were considered beneficial if they could be accessed with “ready-made and integrated tasks”. The notion that creating adequate exercises seemed too time-consuming was also mentioned.

Tyne (2012), who encountered similar positive but skeptical reactions, provides an illuminating summary from his own findings that undoubtedly reflects the attitude of language teachers (who are normally not corpus linguists, bear in mind) more generally:

> “It has been shown that these teachers have quite a pragmatic view of the uses of corpora for language learning: what can *we* get out of it given the many constraints that exist, including the availability of suitable technology (computers or just classroom whiteboard?), student motivation…and predominant teaching-learning methods within the system?” [emphasis original]

All in all, the above research indicates that both student and teacher perceptions vary considerably and this is most likely due to a number of factors. On the students’ end, many have found it to be a rewarding experience, both in terms of language learning and individual growth. However, more technical barriers appear to stand in the way of widespread engagement and acceptance, such as corpus manipulation and relevant content. On the teachers’ side of things, there seems to be an initial enthusiasm that only some are able to transfer over to effective activities, while factors such as insecurity with corpora and time restraints still play a major role in the acceptance of the DDL approach. Because “the decision to incorporate corpora into language teaching lies ultimately with the teacher” (Breyer, 2009), the concerns noted in this section serve to highlight the importance of training teachers so that they feel comfortable and confident enough to work with corpora. Corpora must first be introduced to teachers before students can be expected to take them up. The following section will explore the issue of training in more depth.

1.2.4 Training for Teachers and Students

With the ubiquity of personal computers and the continuous spread of high-speed, wireless internet access, DDL has become more relevant than ever, especially if we consider the tech-savviness of current and future generations. However, in order for DDL to be truly effective, students and teachers must learn how to interact with the corpus, whether it is through concordancing software (e.g. AntConc, Lancsbox, Simple Concordance Program[4], TextSTAT[5], or WordStatix[6]), a web interface, or even paper-based activities with extracted concordance lines. While feeling comfortable with technology is certainly advantageous, consulting a corpus and extracting relevant information is not as simple as looking up a word in a dictionary or doing grammar exercises in a textbook. This is why it is essential for teachers and students to be adequately informed as to what corpora are and what can be done with them if they choose to embark upon corpus-driven learning.

Researchers (Mukherjee, 2006; McCarthy, 2008; Frankenberg-Garcia, 2012) consider a good entry point to be corpus selection and understanding the general nature of what makes a corpus a corpus. Teachers must be made aware of the different corpora available and what characterizes each one. Like any corpus user, basic concepts such as size, language(s), data type, design, and domain must be among the first considerations given by teachers looking to implement corpora in language learning. However, these are frequently overlooked. Frankenberg-Garcia suggests having teachers first try out the same, simple search query in different corpora, which will illustrate the differences more clearly as word frequency and usage will certainly vary among corpora.

Once teachers reach a certain degree of familiarity with the general types of corpora and their characteristics, the next logical aspect to focus on is the formulation of queries. It is important to bear in mind that corpora are not as "intuitive" as an online search engine. Corpora require the user to be as specific as possible. Frankenberg-Garcia (ibid) and Sripicharn (2010) both suggest

[4] http://www.textworld.com/scp/
[5] http://neon.niederlandistik.fu-berlin.de/en/textstat/
[6] https://sites.google.com/site/wordstatix/

first trying with any single word or short strings of words, then adding and removing words to and from the string to see how this affects the number of results. The example Frankenberg-Garcia provides is to begin with *it*, then proceeding to *it was*, then *it was fine*, and so on, while Sripicharn suggests using body parts. Frankenberg-Garcia also suggests trying both inflected and non-inflected words to see how the corpus reacts differently to them. From here, novice users can then begin to experiment with the use of regular expressions as well as specifying part-of-speech (if the corpus is tagged for that). Braun (2006) points out that, "teachers do not make use of traditionally annotated corpora because they are simply not familiar with building complex linguistic queries". Regular expressions allow the user to refine searches and eliminate a lot of extraneous results. This may be one of the more difficult aspects of adequate querying for teachers and students alike due to general unfamiliarity and the steep learning curve when first trying to implement them. However, despite being an initial source of frustration, regular expressions can empower the user in ways simple queries cannot, as will be examined further down in section 3.1.1.3.

Lastly, if teachers are to guide students through activities, they must know how to interpret the data. As many have pointed out (Sripicharn, 2010, for example), concordance lines in and of themselves do not provide any rules. The data must be interpreted, and it must be done in a way that is not counterproductive, such as over-generalization, incorrect rule-formation, or deriving answers from skewed data. Teachers must keep in mind that corpora contain real instances of language use, which means that they will most likely contain non-standard usage and orthography, depending, of course, on their domain. Therefore, teachers should learn how to compare search results in order to get a clearer picture of how the language is being used. To do so, Frankenberg-Garcia suggests a "consciousness-raising exercise" which entails searching for the correct and incorrect spellings of words such as *believe*/*beleive** and *paid*/*payed** in the BNC. While misspellings could be found, the total frequency was negligible (9 and 45, respectively) compared to the correct forms (20,431 and 1,542, respectively). This simple exercise attempts to make teachers aware of the fact that not everything found in a corpus is necessarily indicative of common, standard, or "correct" usage. As for strategies to avoid blanket-generalizations of the results, novice users may be best off working through a DDL task that illustrates the importance of examining the co-texts, which are "(usually small) chunks of text surrounding the examined word or phrase" (Braun, 2006). Frankenberg-Garcia suggests examining prepositions after the word *congratulations* to see how different prepositions affect not only meaning, but also the words that are to follow.

Training for students, then, follows a similar process as training for teachers. However, what differentiates students from teachers is the knowledge of the language, and this should be taken into account when training students. Teachers have far greater intuition with the target language and therefore have a better idea of what types of queries are worth formulating. Kennedy & Miceli (2001), in their study on students learning through "apprenticeship" (which they describe as "promot[ing] learning by example and experience"), state that "whether it makes sense to ask a given question depends to some extent on familiarity with the target language" and that with some learners "there was sometimes insufficient attention to how specific or general a question should be". This highlights once again the importance of the teacher's role in guiding students who have little to no corpus experience. They also observed that "students often did not seem to consciously choose whether to frame their questions in open or closed form", nor did they "seem very concerned that a strategy be efficient. In other words, they did not direct effort at obtaining a workable number of examples". Other issues noted by Kennedy & Miceli in training students included those of observation and interpretation of the data, many of them echoing common issues in teacher training. They concluded that:

> "We recognize that during corpus investigations by language learners, there is considerable room for error due to lack of knowledge of the target language. However, we propose that the development of appropriate research habits – incorporating observation and logical reasoning as well as techniques in corpus searching – could reduce other causes of error to a minimum."

Despite the issues students may have, training need not be too tedious nor time-consuming as "[c]orpora with integrated interfaces for on-line access today may require as little as five minutes' introduction" (Boulton, 2009a). In another publication from that same year Boulton (2009b) writes, "learners can prove quite sophisticated even with complicated tasks such as building and analysing their own corpora with comparatively little training". Finally, we must remember that “corpus skills constitute a learning task in themselves, much in the way that many other subskills of learning do [and]…[o]nce acquired, they facilitate learning greatly and need not be constantly refreshed” (Mauranen, 2004). After some basic training, learners will have a useful skillset that they can continue to develop as they learn.

Section 3.1 below will detail a series of introductory exercises that teachers and students can use to familiarize themselves with LITTERA and its various functions and capabilities. Hopefully through such preparatory tasks, teachers feel comfortable enough to bring corpora into the classroom and guide their students through the same process. Insecurity on the part of the teacher—noted by McCarthy (2008) and Breyer (2009) and a recurring theme in the literature—can and must be overcome through adequate training if corpora are to reach the students.

Given the numerous learning advantages laid forth in this overview, it is possible that if teachers are able to bring DDL into classrooms and make such an approach more “mainstream”, unforeseen solutions will arise to some of the barriers currently keeping DDL locked in the realm of research and away from pedagogical praxis. Good corpus design and relevant material will hopefully encourage teachers to not shy away from DDL, and it is those guiding principles that have led to the creation of the LITTERA corpus, which will be described at length in the following chapter.

2. THE LITTERA CORPUS

Until the turn of the 21st century, speech and multimedia corpora had been far and few. Even spoken corpora—text corpora made up of transcriptions from real instances of speech—had been limited in number and scope for many of the same reasons, mainly time-constraints (particularly in data collection and transcription), cost and technological capacity. However, the last 20 years have seen an uptick in the number of speech and multimedia corpora available, particularly those with pedagogical aims. Despite this, few multimedia corpora have reached a significant size or have a readily accessible web location and interface, save a small handful of specific cases. It is in this context that the LITTERA corpus was conceptualized, not only to fill this gap more generally, but also to create inroads for the study of phonology within DDL. After an initial overview of related work, this chapter will detail the creation and design of the LITTERA corpus and describe at length its composition, features and target user. The chapter will conclude with a discussion on literary language and the role of literature in language learning as this is the domain of the LITTERA corpus.

2.1 Related Work

There are a handful of already existing corpora that are relevant to the development of LITTERA and its implementation in language pedagogy. The corpora discussed here fall into three frequently overlapping categories: Pedagogical Corpora, Parallel Corpora and Speech/Multimedia Corpora.

One of the first corpora to come along and begin collecting speech data was the CHILDES corpus[7], intended to document first language acquisition in children. CHILDES came to fruition in the 1980s and set out to "move beyond the idea of a simple data repository" by creating "shared transcription formats, shared codes, and shared analysis programs" (MacWhinney, 2000),

[7] https://childes.talkbank.org

hence the full name, Child Language Data Exchange System. CHILDES was conceived as a way to develop computerized data that could be accessed by other researchers, much like how many of the corpora to be discussed in this subchapter are freely accessible online, including LITTERA. CHILDES contains numerous corpora from a variety of different languages and can currently be accessed via TalkBank[8], a collection of language data repositories also developed by MacWhinney and of relevance to the present work.

TalkBank is based on the same data sharing principles as CHILDES, but with a much broader scope aimed at providing "community-wide access to naturalistic recordings and transcripts of human and animal communication" (MacWhinney, 2007). It currently contains corpora in 14 different research areas and 34 different languages. The Conversation Banks located within TalkBank are of particular interest to the present work, as they contain recordings of adult conversational speech and can be accessed through the TalkBank homepage.

The most relevant corpora to our purposes here have come about after the turn of the 21st century. Braun's (2005, 2006) ELISA corpus is a pedagogically enriched multimedia corpus with interviews on topics students are likely to encounter in their studies, such as professions or hobbies. The corpus is monolingual (English) and contains audio and video, along with access to activities and other pedagogical material elaborated from the data.

The SCOTS corpus (Anderson et al., 2007) consists of spoken and written texts of the languages of Scotland, along with audio recordings accompanying some of those texts. Among other useful features is a collocate cloud where the user can type a word and generate a word cloud with the collocates from the corpus that displays stronger and weaker collocates by size and text shade.

Another multimedia corpus, SACODEYL (Hoffstaedter & Kohn, 2009) is a compilation of "European teen talk in the context of language education" and includes seven different languages (English, French, German, Italian, Lithuanian, Romanian, and Spanish). Unfortunately, the website where the corpus is found appears to be inactive.

The GeWiss[9] corpus is a multilingual (German, English, Polish and Italian) corpus of spoken academic language in the form of audio recordings and academic communications. It contains

[8] https://talkbank.org

[9] https://gewiss.uni-leipzig.de/index.php?id=about_gewiss&L=1

useful annotations for spoken phenomena such as code-switching, informal speech and overlapping laughter.

FOLK[10] (Schmidt, 2014) is another corpus that contains interactional speech, although in this case entirely in German. It is meant to benefit "a variety of users" and its intended use is not specified, although certainly there is pedagogic potential in such a resource.

The VEIGA[11] corpus (Sotelo Dios & Gómez Guinovart, 2012), located as a sub-corpus within the CLUVI collection (see section 2.2 below for more on CLUVI), is an English-Galician parallel corpus made up of film subtitles, their translations, and audio and video from the films the subtitles were taken from. Sotelo Dios (2016) details the advantages of working with a multimedia corpus in a university translation class through a series of corpus-driven activities and notes that not only were the students able to learn how to use the corpus, but they also found it highly motivating. In addition, Sotelo notes that despite initial struggles to use regular expressions, students reported that this eventually became one of the most satisfying aspects of corpus-driven work.

Finally, perhaps one of the most similar to LITTERA, along with VEIGA, is the TED[12] corpus (Hasebe, 2015), which contains hundreds of videos from TED talks. The corpus can be searched in English, but translations may be selected as well, making this a parallel multimedia corpus. Much like LITTERA, the TED corpus contains scripted speech, although delivered in front of an audience rather than a simple microphone. Aston (2015) used the corpus for his DDL study of phraseological items. The corpus is most in line with what LITTERA aims to achieve, that is, the creation of a large-scale corpus of speech data accompanied by the corresponding text and translation that can be exploited, among other things, for language learning.

Of course, because pedagogical aims often vary, it is practically impossible to have a corpus that encompasses nearly all students' needs in learning a language. As stated previously, which corpus a teacher brings to the classroom and how it is used depends entirely on the learning context and whatever the aims for that lesson may be.

[10] http://agd.ids-mannheim.de/folk.shtml
[11] http://sli.uvigo.gal/CLUVI/index.php?corpus=23&tipo=19&lang=en
[12] https://yohasebe.com/tcse/

2.2 CONCEPTION & COMPOSITION

The LITTERA corpus was originally conceived as a language resource for Spanish-speaking university students of English (Lang & Gómez Guinovart, 2021). It is comprised of 25 fictional literary texts from three different genres (children's literature, short story and novel—both young adult and adult) spanning three centuries (19th, 20th, 21st), with the majority from the 20th century. It is for this reason that the name LITTERA has been chosen for the corpus, as *littera* is the Latin word for *letter* and the origin of the word *literature*. A comprehensive list of the texts included in the corpus can be seen in Table 1 below.

Table 1. Full list of texts in the LITTERA corpus in order of publication year.

Year	*Title & Author*	*Words in English*	*Words in Spanish*	*Total Words*	*Translation Units*	*Genre*
1811	**Sense & Sensibility** Jane Austen	118,503	119,586	238,089	4,681	Novel
1846	**A Tell Tale Heart** Edgar Allen Poe	2,117	1,978	4,159	101	Short Story
1846	**The Cask of Amontillado** Edgar Allen Poe	2,329	2,307	4,653	191	Short Story
1899	**Heart of Darkness** Joseph Conrad	38,765	38,763	77,631	2,252	Novel
1902	**The Hound of the Baskervilles** Sir Arthur Conan Doyle	56,149	56,091	112,240	3,671	Novel
1903	**The Call of the Wild** Jack London	32,035	32,752	64,558	1,612	Novel
1914	**Dubliners** James Joyce	67,978	65,497	133,519	4,531	Novel
1921	**The Mark on the Wall** Virginia Woolf	3,138	3,411	6,549	129	Short Story
1925	**The Great Gatsby** F. Scott Fitzgerald	48,764	50,586	97,667	3,034	Novel
1925	**Mrs. Dalloway** Virginia Woolf	64,249	67,796	131,234	3,266	Novel
1927	**In Another Country** Ernest Hemingway	2,139	2,060	4,199	119	Short Story
1927	**Hills Like White Elephants** Ernest Hemingway	1,459	1,297	2,756	169	Short Story
1945	**The Pearl** John Steinbeck	25,915	26,295	52,210	1,619	Novel
1949	**1984** George Orwell	100,017	95,825	195,842	5,987	Novel
1950	**The Lion, the Witch and the Wardrobe** C.S. Lewis	37,441	36,159	72,676	2,327	Young Adult
1952	**The Old Man and the Sea** Ernest Hemingway	26,657	25,301	51,905	1,830	Novel
1953	**Fahrenheit 451** Ray Bradbury	46,119	46,261	92,399	3,765	Novel

1954	**Lord of the Flies** William Golding	59,128	63,265	124,637	5,747	Novel
1963	**Where the Wild Things Are** Maurice Sendak	339	349	688	33	Children's Lit
1969	**The Very Hungry Caterpillar** Eric Carle	225	218	443	19	Children's Lit.
1981	**The Cathedral** Raymond Carver	6,347	5,637	12,001	679	Short Story
1985	**If You Give a Mouse a Cookie** Laura Numeroff	294	257	548	25	Children's Lit.
1997	**Harry Potter and the Sorcerer's Stone** J.K. Rowling	77,829	77,910	155,276	6,231	Young Adult
2008	**Hunger Games** Suzanne Collins	99,762	99,655	199,621	6,564	Novel
2012	**The Fault in Our Stars** John Green	65,934	65,807	133,109	5,111	Young Adult
	TOTALS	983,632.00	985,063.00	**1,968,695**	**63,693**	

Text selection was primarily based on the English philology curricula at the University of Santiago de Compostela (as of fall of 2016). Although not included in the curricula, children's literature and young adult fiction were added to the corpus in order to provide a wider range of literary styles, registers and voices, especially in the case of young adult fiction as works like *Harry Potter and the Philosopher's Stone*, *The Hunger Games* and *The Fault in Our Stars* contain registers of English generally absent from the so-called *classics*, but of interest to learners of English nonetheless. Furthermore, in the case of young adult literature, these may be books students are already familiar with in Spanish and perhaps willing to take on in English now that they can consult the original texts aligned with their translations.

LITTERA is currently included in two corpora collections. In 2016, LITTERA was originally included as a corpus within the CLUVI corpus collection[13] hosted at the University of Vigo. CLUVI is a nearly 50-million-word collection of parallel corpora with Galician as the primary language, although other language pairs are available (as is the case with LITTERA, English-Spanish). LITTERA can be accessed through the CLUVI home page by following the menu options (*Full Text Search* → *Types of translation* → *Literary translation* → *English-Spanish* [audio icon])[14] or directly via its web address[15].

[13] http://sli.uvigo.gal/CLUVI/index.php?lang=en

[14] After opening the *Full-text search* option, the corpus can also be accessed through the *Language combinations* menu (*Language combinations* → *English-Spanish* → *Literary translation*) or the *Translation languages* menu (*Translation languages* → *English* or *Spanish* → *Literary translation: EN-ES*). The LITTERA corpus is always marked in the menu options with the speaker icon.

[15] http://sli.uvigo.gal/CLUVI/index.php?corpus=24&tipo=1&lang=en

As of June 2020, LITTERA also forms part of the SensoGal[16] corpora collection, which is an open collection of sentence-level aligned parallel corpora, lemmatized and sense-tagged based on WordNet 3.0. It is made up of a selection of corpora from the CLUVI corpus as well as 30 texts from the SemCor Corpus[17]. As will be seen below in section 2.5.2, the SensoGal interface provides options not available through the CLUVI interface that allow the user to carry out searches by semantic concept (according to Galnet) and filter for part of speech.

2.3 TARGET USER PROFILE

The average profile of the target user is a young adult university student between the ages of 18-25 who has had English instruction since about the age of 5 (or earlier in some cases), but most likely still has certain limitations expressing him/herself as more focus throughout primary and secondary education was placed on grammatical forms and vocabulary in lieu of natural speech production and oral comprehension. According to the 2019 EF English Proficiency Index[18], Spain ranks low compared to other EU countries (25th) and received the rating of *moderate proficiency*, placing it 35th worldwide.

In a survey of 28 students from the university's English philology program carried out by the author to form a general student profile (see Appendix A for the full list of survey questions and results), more than half agree with the statement that their primary and secondary school curricula did not focus enough on oral comprehension and expression. In that same survey, more than half say they lack any regular contact with native speakers. Furthermore, 58% (16 students) say that they become nervous when speaking English with a native speaker, and 39% (11 students) also claim to become nervous when speaking English with other Spanish speakers. In addition, 54% (15 students) say they attempt to speak English with a specific accent, from which we can assume that these students are likely very aware of their interlanguage phonology and strive for more "native-like" pronunciation.

The numbers are rather mixed when it comes to time spent abroad in English speaking countries, as well as those who have had supplementary English classes. Only 18% (5 students)

[16] http://sli.uvigo.gal/SensoGal/index.php?corpus=24&tipo=20&lang=en

[17] http://sli.uvigo.gal/SemCor/index_en.php. Semcor is a semantically tagged corpus of English texts created by the WordNet team. In order to be incorporated into SensoGal, the texts were given new semantic labels and any words not already found in Galnet were added to the semantic dictionary (see section 2.5.2 below for more on Galnet).

[18] https://www.ef.com/__/~/media/centralefcom/epi/downloads/full-reports/v9/ef-epi-2019-english.pdf

have spent more than three months at a time in an English speaking country, while the numbers were split fairly evenly between those who have had private English teachers and attended language academies and those who have not.

If one thing is clear from the survey, it is that the students have a clear proclivity for exploiting digital resources, as the vast majority report using streaming websites and social media to practice English. Furthermore, when watching English-language films, 79% (22 students) show some degree of preference for watching them in the original language with English subtitles, while 42% (12 students) also show some preference for Spanish subtitles and 57% (16 students) for no subtitles at all. Yet, the overwhelming majority, 82% (23 students) say that they grew up watching films dubbed into Spanish.

As for their familiarity with corpora, 32% (9 students) claim to know what a corpus is in linguistics, while 15% (4 students) say they have used one in some fashion for learning English. This finding, coupled with students' gravitation towards digital media in language learning, is an encouraging sign for the present work and DDL as a whole, despite the seemingly low numbers. Familiarity with corpora is a necessary first step toward actually implementing them in one's language learning praxis. Still, 28 students is an admittedly small sample size and it is hard to draw any definitive conclusions one way or the other. The idea in administering the 33 question anonymous online survey was to gain a clearer picture of English philology students' perceptions and experiences with the language.

2.4 METHODOLOGY: DEVELOPING THE CORPUS

While the texts were primarily selected based on authors and works found in various syllabi from the English Philology program at the University of Santiago de Compostela, text inclusion was further contingent on the online availability of the original text, a reliable translation in Spanish, and an audiobook. This drastically limited the ability to compile a large-scale literary parallel speech corpus. All translations had to be vetted for overall quality when the translator was unknown, as well as alignment feasibility due to the fact that literary translations often do not mirror the original text, leading to alignment issues. The vetting process for the translations was highly subjective and involved the author reading through random sections of the text to ensure the overall quality of the translation, as blatant translation errors would not be conducive to language learning. In a few cases, texts already aligned by Andres Farkas[19] were used (see the LITTERA homepage for full bibliographical information), although they would be altered during audio segmentation. After text selection, both the original texts and the translations had to be "cleaned up" and put into *.txt* format so they could then be automatically aligned using LF Aligner[20]. Alignments were manually reviewed twice, once right after automatic alignment to check for large gaps and obvious misalignments, and again while editing the audio. This second review had to be much more meticulous as the alignment had to adapt to the audiobooks for reasons that will be explained in the following paragraph. The audiobooks were also vetted for quality since they were taken from publicly available sources, mostly YouTube[21] and Librivox[22]. Professional recordings were always chosen over amateur ones, as the latter had to be carefully examined for mispronunciations and unnatural or *choppy* prosody.

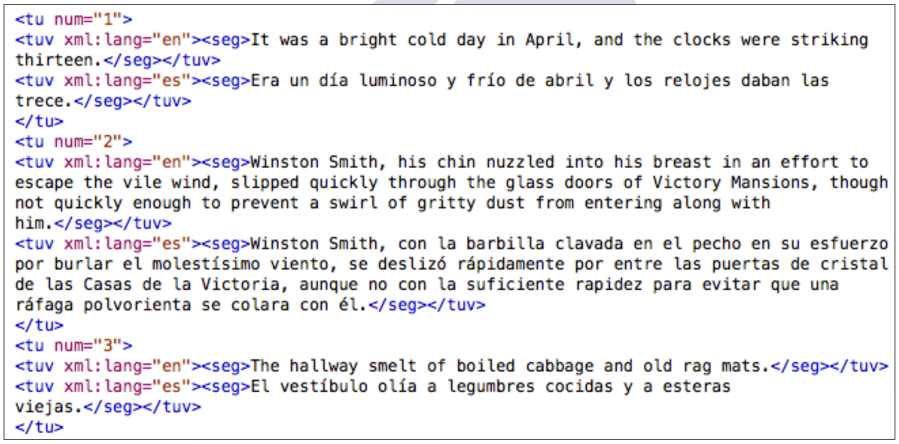
The audio files were manually edited using the open-source audio-editing software Audacity[23]. While it is possible to automatically edit audio files into individual sentences using forced alignment software (e.g. Aeneas Web App[24]) the addition of translations added a third level of segmentation that had to be taken into account, rendering automatic audio segmentation difficult. If one level required the combination of multiple sentences, such as a narrator combining two sentences into a single prosodic unit, the other two levels (text segmentation and text alignment) had to be adjusted as well. While undoubtedly time-consuming, manually editing the audio has ensured a highly reliable alignment with a very limited amount of errors. Those errors remaining are relatively minor and usually involve a misalignment by one TU or a part of a sentence that has been cut off and placed in a neighboring TU.

It should be noted here that the author was responsible for all the aforementioned, including text selection, text retrieval, alignment, audio selection and audio segmentation. One of the author's doctoral supervisors, Xavier Gómez Guinovart, carried out the "back end" operations of designing and managing the corpus's website (along with both the CLUVI and SensoGal

[19] http://farkastranslations.com/
[20] https://sourceforge.net/projects/aligner/
[21] https://www.youtube.com/
[22] http://librivox.org/
[23] https://www.audacityteam.org
[24] https://aeneasweb.org/status

collections more generally) as well as formatting the TMX files generated by the author (Gómez Guinovart, 2019).

Once the texts and audio segments were correctly edited and aligned, the parallel text files were then converted to XML format. Both CLUVI and SensoGal use an adaptation of the TMX format (Savourel, 2005), which is the standard for encoding translation memories. The TMX files for LITTERA are labeled with their corresponding title along with the translation unit number as *tuid* (translation unit identifier), as seen in Figure 1. In the case of SensoGal, semantic annotation is added to the original TMX file. The audio segments are stored as OGG audio files and are labeled numerically with the corresponding book title. This allows users to access the bilingual text pairs and the audio files by searching the LITTERA interface.



```
<tu num="1">
<tuv xml:lang="en"><seg>It was a bright cold day in April, and the clocks were striking
thirteen.</seg></tuv>
<tuv xml:lang="es"><seg>Era un día luminoso y frío de abril y los relojes daban las
trece.</seg></tuv>
</tu>
<tu num="2">
<tuv xml:lang="en"><seg>Winston Smith, his chin nuzzled into his breast in an effort to
escape the vile wind, slipped quickly through the glass doors of Victory Mansions, though
not quickly enough to prevent a swirl of gritty dust from entering along with
him.</seg></tuv>
<tuv xml:lang="es"><seg>Winston Smith, con la barbilla clavada en el pecho en su esfuerzo
por burlar el molestísimo viento, se deslizó rápidamente por entre las puertas de cristal
de las Casas de la Victoria, aunque no con la suficiente rapidez para evitar que una
ráfaga polvorienta se colara con él.</seg></tuv>
</tu>
<tu num="3">
<tuv xml:lang="en"><seg>The hallway smelt of boiled cabbage and old rag mats.</seg></tuv>
<tuv xml:lang="es"><seg>El vestíbulo olía a legumbres cocidas y a esteras
viejas.</seg></tuv>
</tu>
```

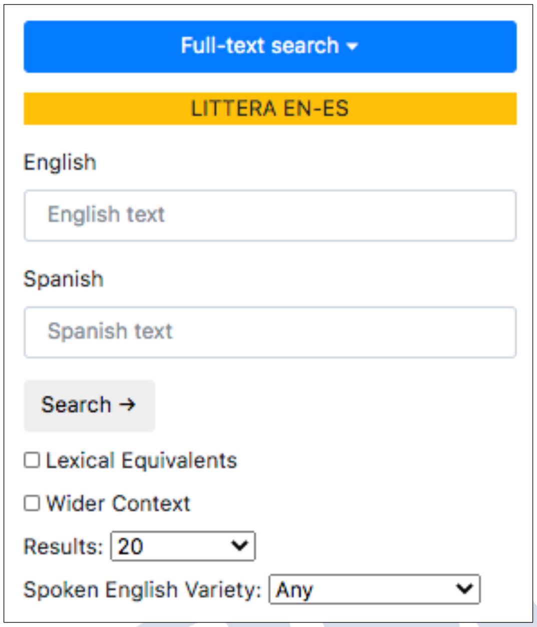
Figure 1. Excerpt from TMX file for George Orwell's *1984*.

2.5 SEARCH QUERIES

The search options available to the user will depend on how the corpus is accessed. As stated previously, LITTERA may be accessed from the CLUVI interface or the SensoGal interface, depending on the type of search and semantic depth desired. In the following sub-chapters 2.5.1 and 2.5.2, the search options available from both interfaces will be described in detail.

2.5.1 Search from CLUVI

Searches through the CLUVI interface may be carried out in English, Spanish or both simultaneously, as can be seen in the screenshot of the search menu in Figure 2.



□ Lexical Equivalents

□ Wider Context

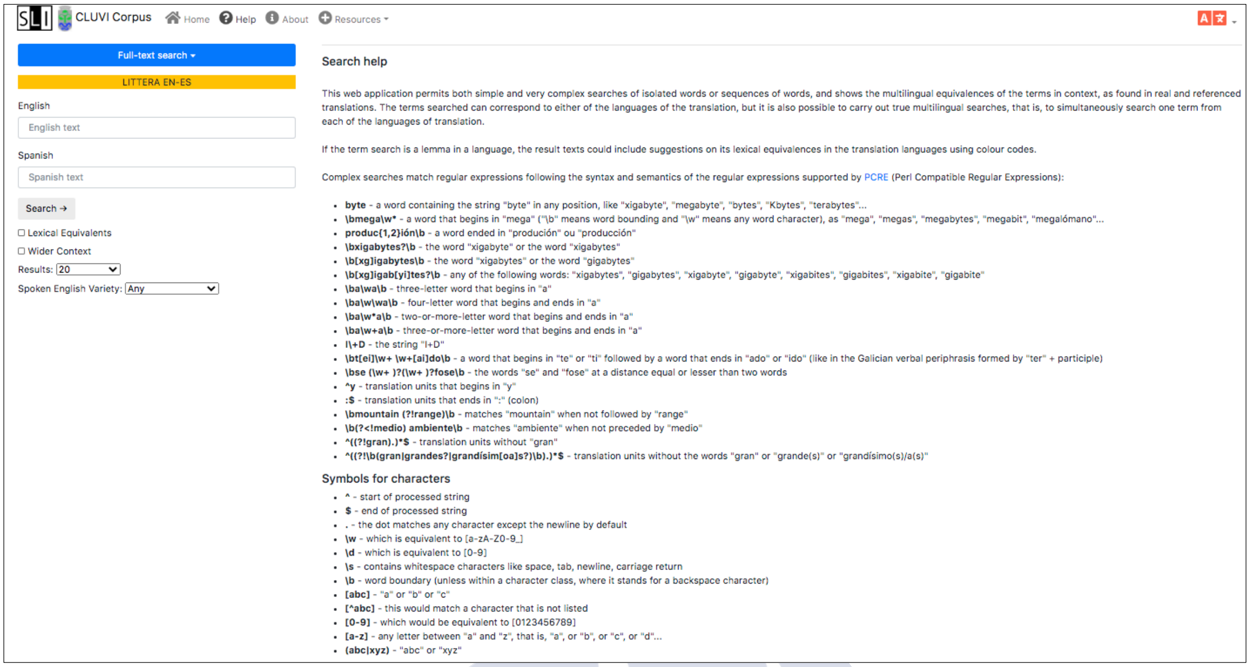
Results: 20

Spoken English Variety: Any

Figure 2. LITTERA search menu in CLUVI.

The corpus allows for simple and complex search queries, and both single and multiple word strings. Complex searches may be carried out through the use of regular expressions, which follow the syntax and semantics of those supported by the PCRE[25] (Pearl Compatible Regular Expressions). A full explanation of the regular expressions available is provided on the Help page (Figure 3), the link to which may be found at the top of the CLUVI interface at all times.

[25] http://www.pcre.org/



□ Lexical Equivalents
□ Wider Context
Results: 20
Spoken English Variety: Any

Search help

This web application permits both simple and very complex searches of isolated words or sequences of words, and shows the multilingual equivalences of the terms in context, as found in real and referenced translations. The terms searched can correspond to either of the languages of the translation, but it is also possible to carry out true multilingual searches, that is, to simultaneously search one term from each of the languages of translation.

If the term search is a lemma in a language, the result texts could include suggestions on its lexical equivalences in the translation languages using colour codes.

Complex searches match regular expressions following the syntax and semantics of the regular expressions supported by PCRE (Perl Compatible Regular Expressions):

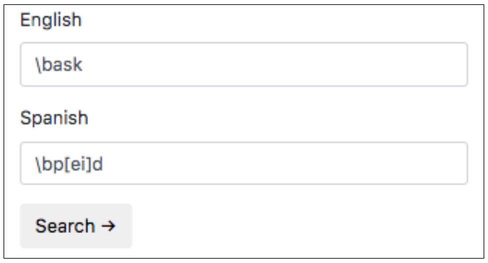
- **byte** - a word containing the string "byte" in any position, like "xigabyte", "megabyte", "bytes", "Kbytes", "terabytes"...
- **\bmega\w*** - a word that begins in "mega" ("\b" means word bounding and "\w" means any word character), as "mega", "megas", "megabytes", "megabit", "megalómano"...
- **produc{1,2}ión\b** - a word ended in "produción" ou "producción"
- **\bxigabytes?\b** - the word "xigabyte" or the word "xigabytes"
- **\b[xg]igabytes\b** - the word "xigabytes" or the word "gigabytes"
- **\b[xg]igab[yi]tes?\b** - any of the following words: "xigabytes", "gigabytes", "xigabyte", "gigabyte", "xigabites", "gigabites", "xigabite", "gigabite"
- **\ba\wa\b** - three-letter word that begins in "a"
- **\ba\w\wa\b** - four-letter word that begins and ends in "a"
- **\ba\w*a\b** - two-or-more-letter word that begins and ends in "a"
- **\ba\w+a\b** - three-or-more-letter word that begins and ends in "a"
- **I\+D** - the string "I+D"
- **\bt[ei]\w+ \w+[ai]do\b** - a word that begins in "te" or "ti" followed by a word that ends in "ado" or "ido" (like in the Galician verbal periphrasis formed by "ter" + participle)
- **\bse (\w+)?(\w+)?fose\b** - the words "se" and "fose" at a distance equal or lesser than two words
- **^y** - translation units that begins in "y"
- **:$** - translation units that ends in ":" (colon)
- **\bmountain (?!range)\b** - matches "mountain" when not followed by "range"
- **\b(?<!medio) ambiente\b** - matches "ambiente" when not preceded by "medio"
- **^((?!gran).)*$** - translation units without "gran"
- **^((?!\b(gran|grandes?|grandísim[oa]s?)\b).)*$** - translation units without the words "gran" or "grande(s)" or "grandísimo(s)/a(s)"

Symbols for characters

- **^** - start of processed string
- **$** - end of processed string
- **.** - the dot matches any character except the newline by default
- **\w** - which is equivalent to [a-zA-Z0-9_]
- **\d** - which is equivalent to [0-9]
- **\s** - contains whitespace characters like space, tab, newline, carriage return
- **\b** - word boundary (unless within a character class, where it stands for a backspace character)
- **[abc]** - "a" or "b" or "c"
- **[^abc]** - this would match a character that is not listed
- **[0-9]** - which would be equivalent to [0123456789]
- **[a-z]** - any letter between "a" and "z", that is, "a", or "b", or "c", or "d"...
- **(abc|xyz)** - "abc" or "xyz"

Figure 3. CLUVI Help page.

Although their range of capabilities may require some time and practice to fully grasp, regular expressions provide the user with the option to search multiple items at a time and compare the results on the same page. An example of this would be if the user wanted to see all instances of *ask* when translated as *pedir*. Because *pedir* is a stem-changing verb, it is necessary to use some type of regular expression in order to see all instances within a single query. The simplest option would be to use brackets, as in Figure 4.

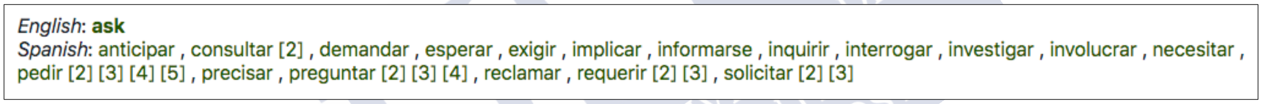


English
\bask
Spanish
\bp[ei]d

Figure 4. Search in CLUVI using regular expressions.

This query tells the corpus to find all instances of any character contained within the brackets, in this case -*e*- or -*i*-, which ensures that the stems *ped*- or *pid*- are included in the results. *\b* marks a word boundary and prevents the corpus from including words such as *despedir*. The use of regular expressions brings us back to the previous discussion in section 1.2.4 regarding training, both for students and teachers. Although it is certainly possible to learn how to use regular expressions on one's own, it should not be taken for granted that students (or educators) will be able to do so without some form of guidance. In chapter 3, some examples of practical training exercises are presented and discussed to familiarize the user with the most common and useful regular expressions available in CLUVI.

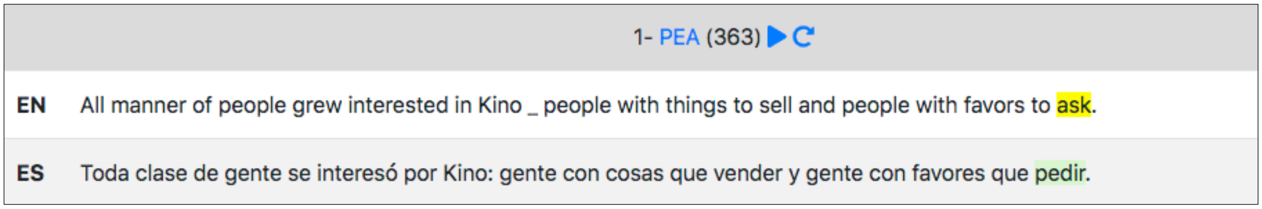
Further search options are available to the user below the text boxes. The Lexical Equivalents option, which is ticked by default, allows the user to search by lemma and highlights the equivalent word in the translation. A list of equivalents also appears at the top of the page based on the available "sense" definitions in the semantic dictionary Galnet (see 2.5.2 below for more on Galnet). Thus, if we tick this option and enter *ask* in the English search box (leaving the Spanish search box empty), we will be given the dictionary list as seen in Figure 5 below.



English: **ask**
Spanish: anticipar , consultar [2] , demandar , esperar , exigir , implicar , informarse , inquirir , interrogar , investigar , involucrar , necesitar , pedir [2] [3] [4] [5] , precisar , preguntar [2] [3] [4] , reclamar , requerir [2] [3] , solicitar [2] [3]

Figure 5. Galnet entry for *ask* which appears at the top of the results page.

It is important to note that the Lexical Equivalents option only responds to lemmas. Therefore, if *asked* is entered instead of *ask*, no definitions will be provided at the top of the page, nor will any words be highlighted in green in the translation as would normally occur, as in Figure 6 below.



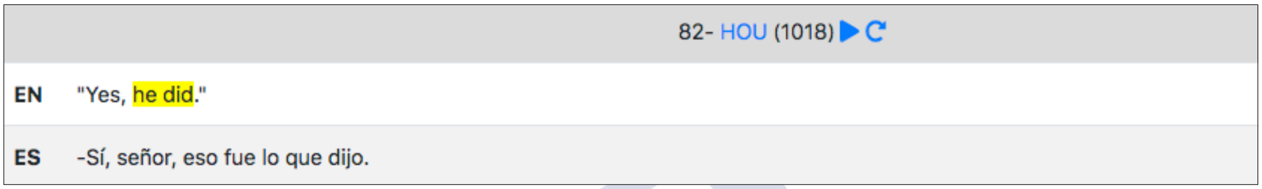
1- PEA (363) ▶ ↻

EN All manner of people grew interested in Kino _ people with things to sell and people with favors to ask.

ES Toda clase de gente se interesó por Kino: gente con cosas que vender y gente con favores que pedir.

Figure 6. The lexical equivalent *pedir* highlighted in green in the translation.

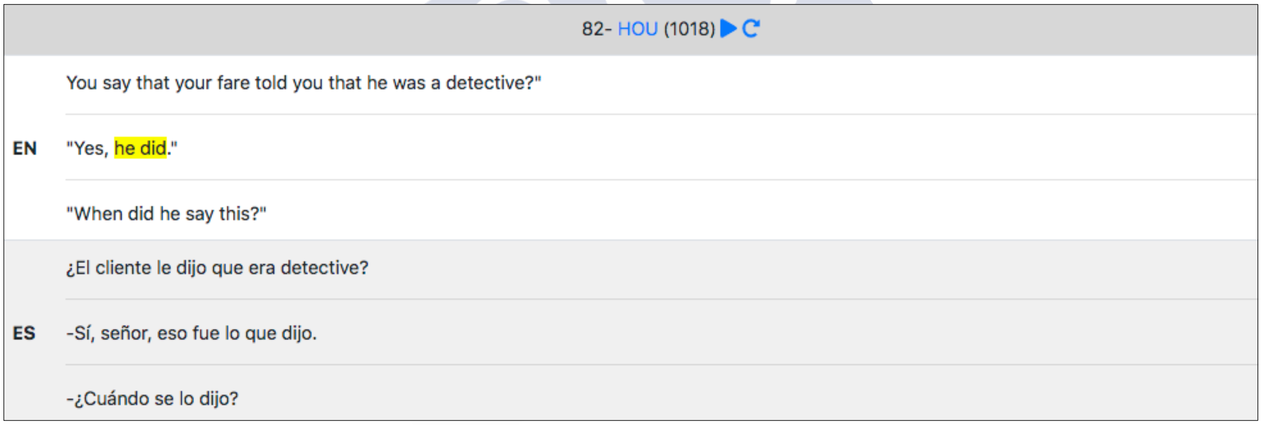
The next option located just below the Lexical Equivalents option is the Wider Context option, un-ticked by default. By ticking this box, any search results will display three translation units instead of one, with the searched item always appearing in the central TU. Normally when presented as search results, TU's are stripped of all surrounding context. This can lead to ambiguity in both the original and in the translation. Therefore, it is necessary to consult the surrounding TU's in order to contextualize the search results and clear up any ambiguity. Figures 7 & 8 below show how context allows the user to better understand the use of *he did* and its translation.



82- HOU (1018) ▶ ↻

EN	"Yes, he did."
ES	-Sí, señor, eso fue lo que dijo.

Figure 7. Translation unit without wider context.



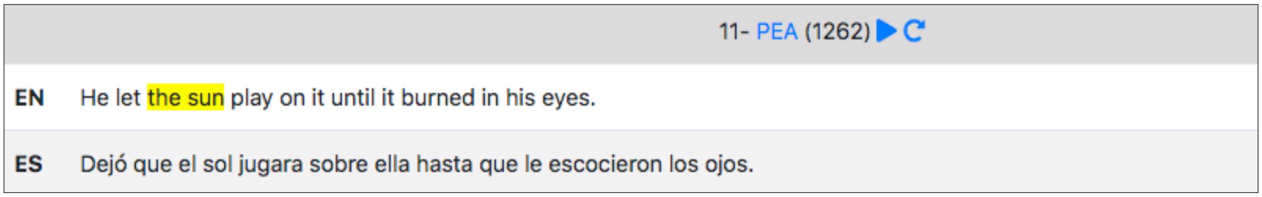
82- HOU (1018) ▶ ↻

	You say that your fare told you that he was a detective?"
EN	"Yes, he did."
	"When did he say this?"
	¿El cliente le dijo que era detective?
ES	-Sí, señor, eso fue lo que dijo.
	-¿Cuándo se lo dijo?

Figure 8. Translation unit displaying the wider context.

Because it is possible in English to respond to a question using only an auxiliary verb, in this case *did*, we need more context to understand the real meaning of *yes, he did* and why it was translated as *sí, señor, eso fue lo que dijo*. Context may also help resolve any confusion as a result of an alignment error that might have been overlooked when editing the audio files.

Another instance in which context helps to clear up confusion can be seen in Figures 9 & 10. In this case, *it* is translated as *ella* and the context disambiguates the use of the two pronouns, which refer to "the pearl" and "la perla", respectively.

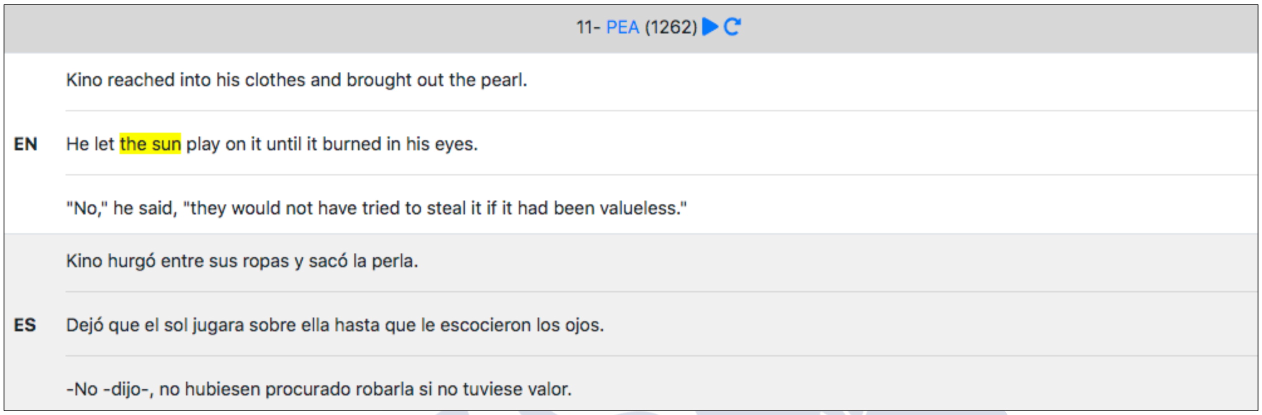


11- PEA (1262)

EN He let the sun play on it until it burned in his eyes.

ES Dejó que el sol jugara sobre ella hasta que le escocieron los ojos.

Figure 9. Translation unit in which there is ambiguity with the pronouns *it* and *ella*.



11- PEA (1262)

Kino reached into his clothes and brought out the pearl.

EN He let the sun play on it until it burned in his eyes.

"No," he said, "they would not have tried to steal it if it had been valueless."

Kino hurgó entre sus ropas y sacó la perla.

ES Dejó que el sol jugara sobre ella hasta que le escocieron los ojos.

-No -dijo-, no hubiesen procurado robarla si no tuviese valor.

Figure 10. Translation unit displaying the wider context to disambiguate the pronouns *it* and *ella*.

Along with the aforementioned, there are still two further search options: Results and Spoken English Variety, both in the form of drop-down menus. The Results option, as seen in Figure 11, controls how many TU's appear on the page as well as how they are distributed. The corpus is automatically set to return the first 20 results. This may not be the most desired setting, however, in the case of high-frequency items due to the fact that the first twenty instances will most likely come from the same text, *The Pearl*. In order to display a wider variety of results, the user may limit the number of results per text to 1, 5, 10 or 15. Variety in terms of texts will also mean variety in terms of narrators and the possibility to hear different voices, rhythms and accents. Other options include viewing the first 50 or 100 TU's, or all of the results (with a limit of 1,500). Finally, the Spoken English Variety option, seen in Figure 12, allows the user to choose between British and North American English.

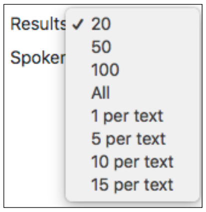


Figure 11. Results menu.

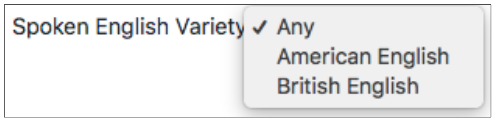
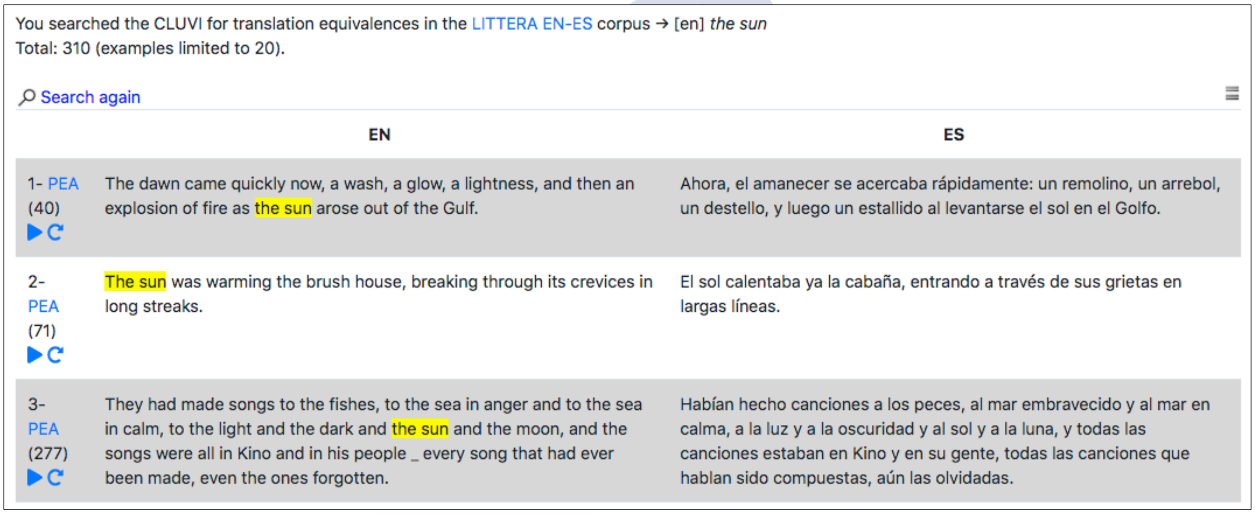


Figure 12. Spoken English Variety menu.

After a search has been carried out, the user may change the presentation format from the default stacked horizontal view, as seen in Figures 6-10 above, to that of two vertical columns, one for each language, which allows for a side-by-side comparison of the TU's, as in Figure 13 below.



You searched the CLUVI for translation equivalences in the LITTERA EN-ES corpus → [en] *the sun*
Total: 310 (examples limited to 20).

	EN	ES
1- PEA (40) ▶ ↻	The dawn came quickly now, a wash, a glow, a lightness, and then an explosion of fire as the sun arose out of the Gulf.	Ahora, el amanecer se acercaba rápidamente: un remolino, un arrebol, un destello, y luego un estallido al levantarse el sol en el Golfo.
2- PEA (71) ▶ ↻	The sun was warming the brush house, breaking through its crevices in long streaks.	El sol calentaba ya la cabaña, entrando a través de sus grietas en largas líneas.
3- PEA (277) ▶ ↻	They had made songs to the fishes, to the sea in anger and to the sea in calm, to the light and the dark and the sun and the moon, and the songs were all in Kino and in his people _ every song that had ever been made, even the ones forgotten.	Habían hecho canciones a los peces, al mar embravecido y al mar en calma, a la luz y a la oscuridad y al sol y a la luna, y todas las canciones estaban en Kino y en su gente, todas las canciones que habían sido compuestas, aún las olvidadas.

Figure 13. Search results in vertical view.

Accompanying each TU is the three-letter code specifying which text it comes from, the TU's number within that text, a play option and a two second rewind option, as in Figure 14. Unlike many other contemporary speech corpora (e.g. TED corpus & SCOTS corpus from 2.1 above), LITTERA provides the option to play the audio right from the same page as the results. In many other speech corpora, the user is redirected to a specific page with the audio/video file in question. This is more time-consuming, especially if the user wants to explore a variety of instances of speech.

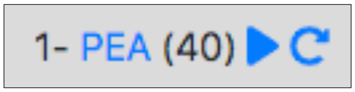


Figure 14. Three-letter code, TU number and audio options.

Lastly, it is also possible to search within a single text via the Audiobooks option, available in the CLUVI menu, as seen in Figure 15. Searching within an individual work allows the student to operate within a smaller corpus, which some researchers have argued as beneficial for language learners (Thompson & Tribble, 2001; Braun, 2005). Limiting corpus examination to one text may make it easier for students to *authenticate* the text in Widdowson's sense of the word, especially if it is a text that they have had to read for class or are simply interested in. This is why the authentication process may be easier for texts within LITTERA than other corpora; these texts are immediately relevant to students needs as many of them form part of the English curriculum or popular culture. Furthermore, searching within a single work allows for a stylistic analysis of a given text, expanding the corpus's potential applications beyond the language classroom and into the realm of literary analysis. Students can analyze patterns within a text, which can shed light upon an author's style and idiosyncratic use of language. For more on literary analysis through LITTERA, see section 2.7 below.

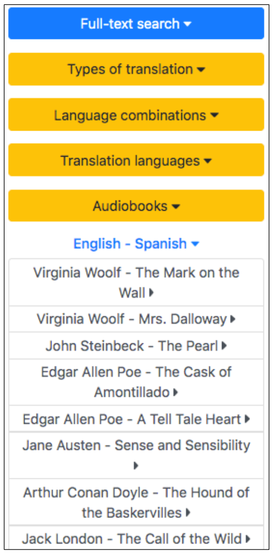

Full-text search ▾
Types of translation ▾
Language combinations ▾
Translation languages ▾
Audiobooks ▾
English - Spanish ▾
Virginia Woolf - The Mark on the Wall ▸
Virginia Woolf - Mrs. Dalloway ▸
John Steinbeck - The Pearl ▸
Edgar Allen Poe - The Cask of Amontillado ▸
Edgar Allen Poe - A Tell Tale Heart ▸
Jane Austen - Sense and Sensibility ▸
Arthur Conan Doyle - The Hound of the Baskervilles ▸
Jack London - The Call of the Wild ▸

Figure 15. Audiobooks menu.

2.5.2 Search from SensoGal

Many of the same features from the CLUVI search menu are available in the SensoGal menu as well. However, because the corpora included in SensoGal have been semantically tagged, there are a number of additional search options available. Figure 16 provides an image of the search menu.

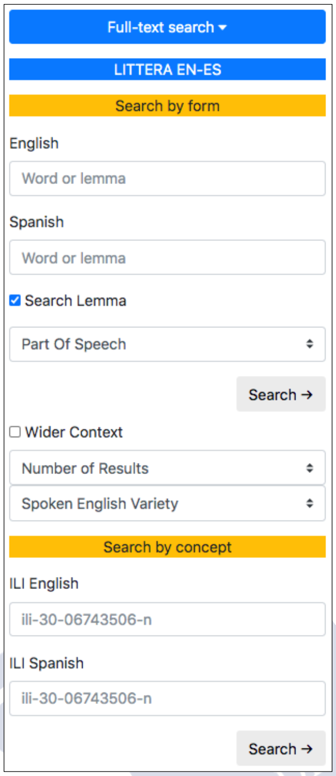
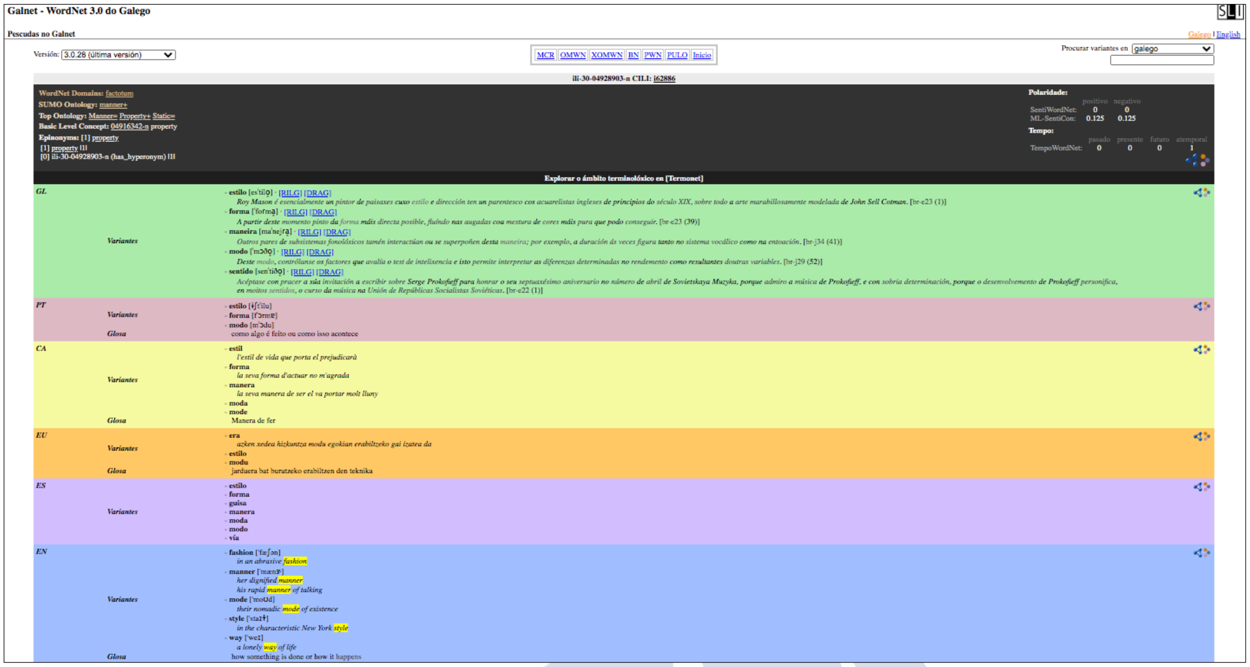


Figure 16. LITTERA search menu in SensoGal.

The fact that SensoGal is a semantically annotated corpus (i.e. sense-tagged) means that many of the lexical words in the corpus (i.e. nouns, verbs, adjectives and adverbs) are assigned an Inter-linguistic Index or ILI code based on WordNet 3.0. The ILI codes are a type of ID-number for a specific concept, or sense, and are index codes for Galnet, the Galician-based WordNet. For example, the ILI *ili-30-04928903-n*[26] refers to the concept "how something is done or how it happens". The words listed in English for this concept include *fashion*, *manner*, *mode*, *style* and *way*, and the words listed in Spanish include *estilo*, *forma*, *guisa*, *manera*, *moda*, *modo* and *vía*. Other languages with entries for this concept include Galician, Portuguese, Catalan, Basque, German, Latin, Italian and French.

[26] For more information on ILI codes, see section 3.1.2.2, Search by Concept



Galnet - WordNet 3.0 do Galego

Pescudas no Galnet

Galego | English

Versión: 3.0.28 (última versión)

MCR | OMWN | XOMWN | BN | PWN | PULO | Inicio

Procurar variantes en galego

ili-30-04928903-n CILI: i62886

WordNet Domains: factotum
SUMO Ontology: manner+
Top Ontology: Manner= Property+ Static=
Basic Level Concept: 04916342-n property
Epinonyms: [1] property
[1] property |||
[0] ili-30-04928903-n (has_hyperonym) |||

Polaridade:

	positivo	negativo
SentiWordNet:	0	0
ML-SentiCon:	0.125	0.125

Tempo:

	pasado	presente	futuro	atemporal
TempoWordNet:	0	0	0	1

Explorar o ámbito terminolóxico en [Termonet]

GL

Variantes
- **estilo** [es'tilo̞] · [RILG] [DRAG]
  Roy Mason é esencialmente un pintor de paisaxes cuxo estilo e dirección ten un parentesco cos acuarelistas ingleses de principios do século XIX, sobre todo a arte marabillosamente modelada de John Sell Cotman. [br-e23 (1)]
- **forma** ['foɾma̞] · [RILG] [DRAG]
  A partir deste momento pinto da forma máis directa posible, fluíndo nas augadas coa mestura de cores máis pura que podo conseguir. [br-e23 (39)]
- **maneira** [ma'nejɾa̞] · [RILG] [DRAG]
  Outros pares de subsistemas fonolóxicos tamén interactúan ou se superpoñen desta maneira; por exemplo, a duración ás veces figura tanto no sistema vocálico como na entoación. [br-j34 (41)]
- **modo** ['mɔðo̞] · [RILG] [DRAG]
  Deste modo, contrólanse os factores que avalía o test de intelixencia e isto permite interpretar as diferenzas determinadas no rendemento como resultantes doutras variables. [br-j29 (52)]
- **sentido** [sen'tiðo̞] · [RILG] [DRAG]
  Acéptase con pracer a súa invitación a escribir sobre Serge Prokofieff para honrar o seu septuaxésimo aniversario no número de abril de Sovietskaya Muzyka, porque admiro a música de Prokofieff, e con sobria determinación, porque o desenvolvemento de Prokofieff personifica, en moitos sentidos, o curso da música na Unión de Repúblicas Socialistas Soviéticas. [br-e22 (1)]

PT

Variantes
- **estilo** [ɨʃt'ilu]
- **forma** [f'ɔrmɐ]
- **modo** [m'ɔdu]

Glosa como algo é feito ou como isso acontece

CA

Variantes
- **estil**
  l'estil de vida que porta el prejudicarà
- **forma**
  la seva forma d'actuar no m'agrada
- **manera**
  la seva manera de ser el va portar molt lluny
- **moda**
- **mode**

Glosa Manera de fer

EU

Variantes
- **era**
  azken xedea hizkuntza modu egokian erabiltzeko gai izatea da
- **estilo**
- **modu**

Glosa jarduera bat burutzeko erabiltzen den teknika

ES

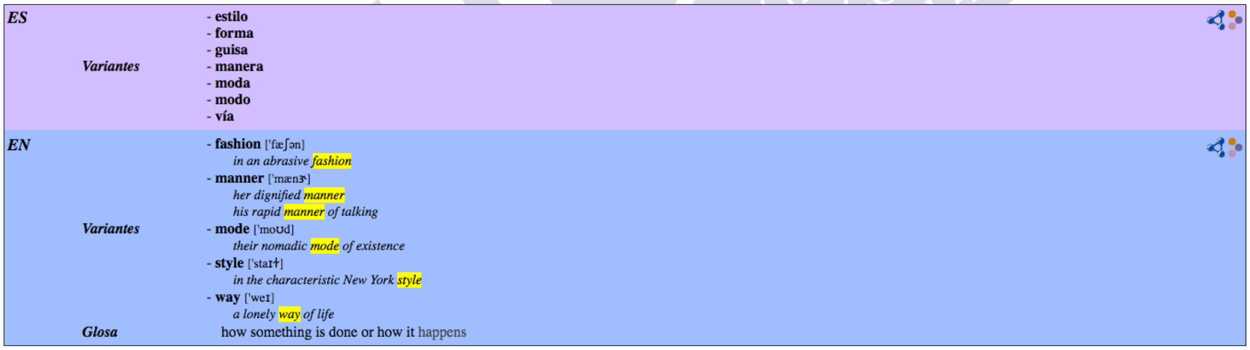
Variantes
- **estilo**
- **forma**
- **guisa**
- **manera**
- **moda**
- **modo**
- **vía**

EN

Variantes
- **fashion** ['fæʃən]
  in an abrasive fashion
- **manner** ['mænɝ]
  her dignified manner
  his rapid manner of talking
- **mode** ['moʊd]
  their nomadic mode of existence
- **style** ['staɪɫ]
  in the characteristic New York style
- **way** ['weɪ]
  a lonely way of life

Glosa how something is done or how it happens

Figure 17. Galnet page for ILI 30-04928903-n displaying Galician, Portuguese, Catalan, Basque, Spanish and English entries.



ES

Variantes
- **estilo**
- **forma**
- **guisa**
- **manera**
- **moda**
- **modo**
- **vía**

EN

Variantes
- **fashion** ['fæʃən]
  in an abrasive fashion
- **manner** ['mænɝ]
  her dignified manner
  his rapid manner of talking
- **mode** ['moʊd]
  their nomadic mode of existence
- **style** ['staɪɫ]
  in the characteristic New York style
- **way** ['weɪ]
  a lonely way of life

Glosa how something is done or how it happens

Figure 18. Spanish and English entries in Galnet for ILI 30-04928903-n.

Searches in SensoGal may be carried out by form (word or lemma) or concept. Unlike CLUVI, SensoGal does not allow for the use of regular expressions. To search by word, the user only has to un-tick the *Search Lemma* option available just below the search boxes. A search for *take* by word will return all instances of *take* and no other variants. When searching by lemma, a search for *take* returns all variations, including *take*, *took*, *taking*, *taken*, etc.

When beginning a search query by word or lemma, autocomplete options appear in a drop down list from the search box for both the English and Spanish options. The predictive list only displays words contained within the corpus and are presented in alphabetical order. Figure 19 shows the autocomplete options when *an* is entered into the English search box.

Also available when searching by word or lemma is the ability to filter for part of speech. Because SensoGal is tagged semantically, this option only includes lexical words and not grammatical words such as prepositions and articles.

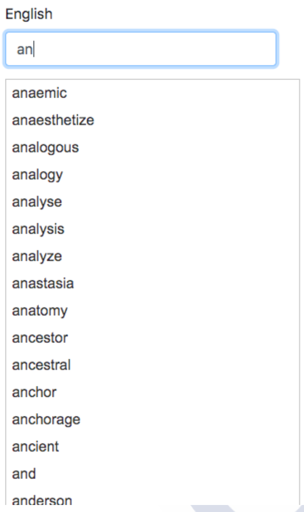


Figure 19. Autocomplete dropdown menu for *an*.

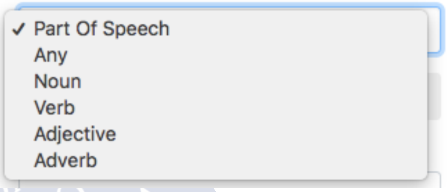
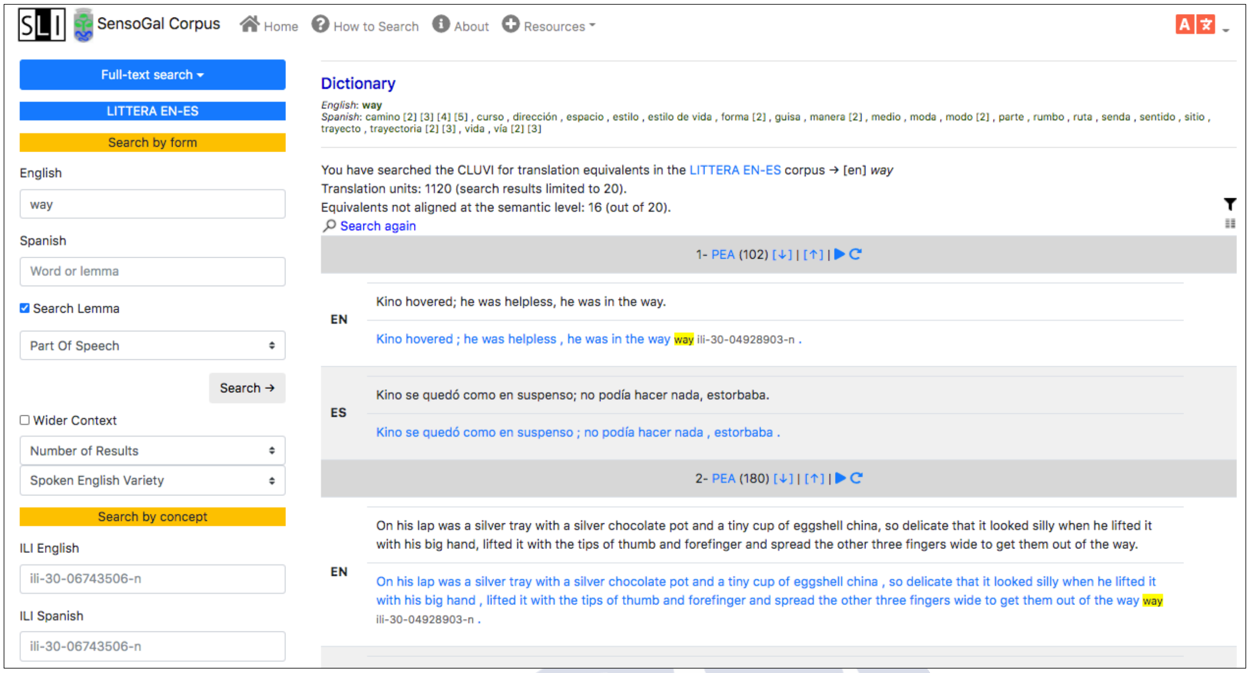


Figure 20. Part of Speech dropdown menu.

The search results display in SensoGal varies slightly from that of CLUVI. While still the same basic format (along with the same vertical and horizontal options as before), both the English and Spanish sentences in the translation units are duplicated in blue underneath the original sentences with the corresponding semantic tag, as seen in Figure 21 below:



English: way

Spanish: Word or lemma

☑ Search Lemma

Part Of Speech

☐ Wider Context

Number of Results

Spoken English Variety

ILI English: ili-30-06743506-n

ILI Spanish: ili-30-06743506-n

Dictionary

English: **way**
Spanish: camino [2] [3] [4] [5] , curso , dirección , espacio , estilo , estilo de vida , forma [2] , guisa , manera [2] , medio , moda , modo [2] , parte , rumbo , ruta , senda , sentido , sitio , trayecto , trayectoria [2] [3] , vida , vía [2] [3]

You have searched the CLUVI for translation equivalents in the LITTERA EN-ES corpus → [en] *way*
Translation units: 1120 (search results limited to 20).
Equivalents not aligned at the semantic level: 16 (out of 20).
Search again

1- PEA (102) [↓] | [↑] | ▶ ↻

EN
Kino hovered; he was helpless, he was in the way.
Kino hovered ; he was helpless , he was in the way way ili-30-04928903-n .

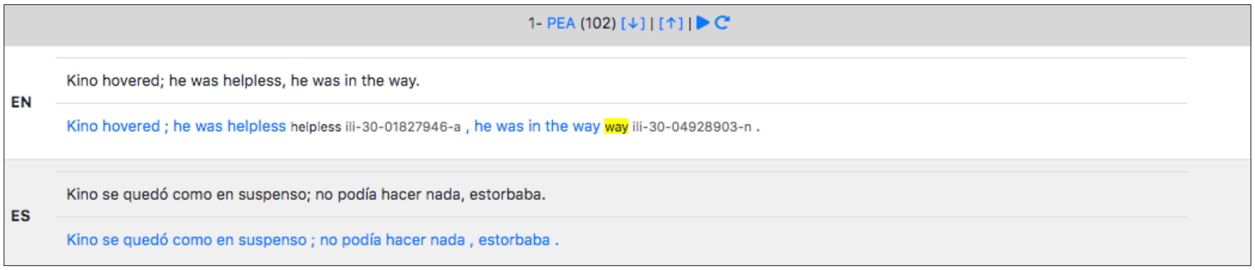
ES
Kino se quedó como en suspenso; no podía hacer nada, estorbaba.
Kino se quedó como en suspenso ; no podía hacer nada , estorbaba .

2- PEA (180) [↓] | [↑] | ▶ ↻

EN
On his lap was a silver tray with a silver chocolate pot and a tiny cup of eggshell china, so delicate that it looked silly when he lifted it with his big hand, lifted it with the tips of thumb and forefinger and spread the other three fingers wide to get them out of the way.
On his lap was a silver tray with a silver chocolate pot and a tiny cup of eggshell china , so delicate that it looked silly when he lifted it with his big hand , lifted it with the tips of thumb and forefinger and spread the other three fingers wide to get them out of the way way ili-30-04928903-n .

Figure 21. View of search results for *way* in SensoGal by lemma.

Note that the ILI code is visible next to the queried word *way*. The ILI codes for any other lexical words (as long as they are found in Galnet and correctly tagged in the corpus) can be viewed by clicking on the word (as in Figure 22 below), or all ILI codes can be viewed at once with the downward arrow option located between the results index number and the play option when viewed horizontally (Figure 23 below). The upward arrow hides the ILI codes again. All available ILI codes link to the word's entry in Galnet where further sense relationships can be explored between a variety of languages, depending on the semantic concept in question.

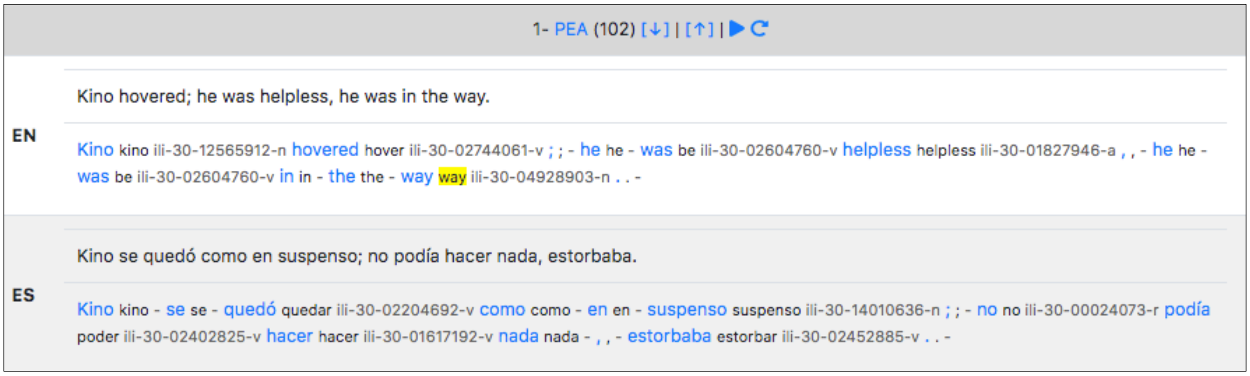


1- PEA (102) [↓] | [↑] | ▶ ↻

EN
Kino hovered; he was helpless, he was in the way.
Kino hovered ; he was helpless helpless ili-30-01827946-a , he was in the way way ili-30-04928903-n .

ES
Kino se quedó como en suspenso; no podía hacer nada, estorbaba.
Kino se quedó como en suspenso ; no podía hacer nada , estorbaba .

Figure 22. View of the ILI's for *way* and *helpless*, the latter displayed by clicking on the word.



1- PEA (102) [↓] | [↑] | ▶ ↻

EN

Kino hovered; he was helpless, he was in the way.

Kino kino ili-30-12565912-n hovered hover ili-30-02744061-v ; ; - he he - was be ili-30-02604760-v helpless helpless ili-30-01827946-a , , - he he - was be ili-30-02604760-v in in - the the - way way ili-30-04928903-n . . -

ES

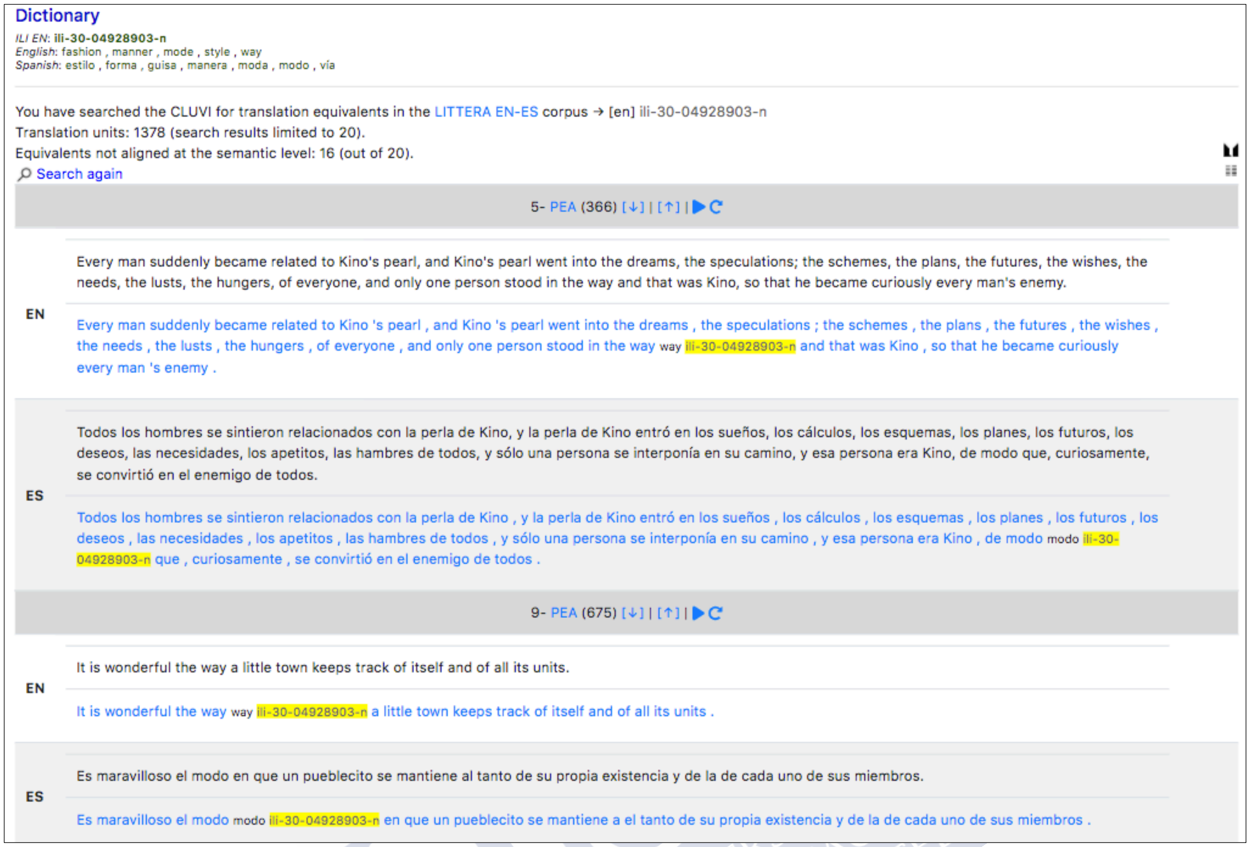
Kino se quedó como en suspenso; no podía hacer nada, estorbaba.

Kino kino - se se - quedó quedar ili-30-02204692-v como como - en en - suspenso suspenso ili-30-14010636-n ; ; - no no ili-30-00024073-r podía poder ili-30-02402825-v hacer hacer ili-30-01617192-v nada nada - , , - estorbaba estorbar ili-30-02452885-v . . -

Figure 23. View of all ILI's available in the TU, displayed by clicking on the downward arrow.

The ILI codes are used in order to search by concept. This can be done in English or Spanish or both simultaneously. If we search for the same ILI code from Figures 21-23 above for the concept "how something is done or how it happens", *ili-30-04928903-n*, many of the same results will appear, particularly those for *way*, although other words corresponding to the concept will be returned, such as *manner*.

Similar to searches via CLUVI, additional information is provided at the top of the page above the search results when searching by concept, as seen in Figure 24 below. First is the Galnet dictionary result, which lists the ILI reference, the English words for the concept, and finally the Spanish words (the order of the languages is flipped when searching in Spanish only). Additional information about the search can be found just below the Galnet dictionary entries, including the search query, the number of total TU's and the number of TU's displayed. However, SensoGal also provides information on the number of cases in which there is no corresponding word to that concept in the other language, stated as *Equivalents not aligned at semantic level: 16 (out of 20)*. These non-alignments can be removed from view by using the funnel icon in the upper right above the search results.



Dictionary

ILI EN: **ili-30-04928903-n**
English: fashion , manner , mode , style , way
Spanish: estilo , forma , guisa , manera , moda , modo , vía

You have searched the CLUVI for translation equivalents in the LITTERA EN-ES corpus → [en] ili-30-04928903-n
Translation units: 1378 (search results limited to 20).
Equivalents not aligned at the semantic level: 16 (out of 20).
Search again

5- PEA (366) [↓] | [↑] | ▶ C

EN

Every man suddenly became related to Kino's pearl, and Kino's pearl went into the dreams, the speculations; the schemes, the plans, the futures, the wishes, the needs, the lusts, the hungers, of everyone, and only one person stood in the way and that was Kino, so that he became curiously every man's enemy.

Every man suddenly became related to Kino 's pearl , and Kino 's pearl went into the dreams , the speculations ; the schemes , the plans , the futures , the wishes , the needs , the lusts , the hungers , of everyone , and only one person stood in the way way ili-30-04928903-n and that was Kino , so that he became curiously every man 's enemy .

ES

Todos los hombres se sintieron relacionados con la perla de Kino, y la perla de Kino entró en los sueños, los cálculos, los esquemas, los planes, los futuros, los deseos, las necesidades, los apetitos, las hambres de todos, y sólo una persona se interponía en su camino, y esa persona era Kino, de modo que, curiosamente, se convirtió en el enemigo de todos.

Todos los hombres se sintieron relacionados con la perla de Kino , y la perla de Kino entró en los sueños , los cálculos , los esquemas , los planes , los futuros , los deseos , las necesidades , los apetitos , las hambres de todos , y sólo una persona se interponía en su camino , y esa persona era Kino , de modo modo ili-30-04928903-n que , curiosamente , se convirtió en el enemigo de todos .

9- PEA (675) [↓] | [↑] | ▶ C

EN

It is wonderful the way a little town keeps track of itself and of all its units.

It is wonderful the way way ili-30-04928903-n a little town keeps track of itself and of all its units .

ES

Es maravilloso el modo en que un pueblecito se mantiene al tanto de su propia existencia y de la de cada uno de sus miembros.

Es maravilloso el modo modo ili-30-04928903-n en que un pueblecito se mantiene a el tanto de su propia existencia y de la de cada uno de sus miembros .

Figure 24. SensoGal search results for the concept tagged as *ili-30-04928903-n*; showing only equivalents aligned at the semantic level (via the funnel icon in the upper right).

Note that the ILI code is highlighted in yellow in both the English and Spanish annotated sentences when the filter (the funnel icon) is applied.

The decision regarding which interface to use will depend on the aims of students and teachers. CLUVI provides the option to search by using regular expressions, which allows the user to tweak a given search to obtain more flexible results, while SensoGal provides tools for a more in-depth semantic analysis of the corpus data.

2.6 Representativeness & LITTERA

When discussing representativeness in the context of the LITTERA corpus, it is important to first make the distinction, as Braun (2005) does, between corpora for linguistic research and pedagogical corpora. Because the aims of these two types of corpora are different, they cannot be expected to follow the same methodology of corpus compilation, as pedagogical corpora must be compiled in accordance to learners' needs.

LITTERA is intended to be representative of two things: 1) works of literature found in the English philology curricula and of general interest to students due to their cultural relevance, e.g. *Harry Potter and the Philosopher's Stone*, and 2) segmental and suprasegmental aspects of English spoken by native speakers. The corpus is not intended to represent spontaneous English speech. While most of the texts in the corpus contain dialogue, it is still scripted speech and depends on the narrator to "bring it to life" and make it feel as natural as possible. For this reason, the focus of the present work is on the phonetic data available in the corpus at the segmental and suprasegmental level as these are the nuts and bolts of spoken English that vary only in minor ways between scripted and spontaneous speech. More conversational aspects of speech such as intonation are not examined here, as it is the author's opinion that this corpus is not suitable for such research. That is not to say that insights into intonation cannot be gained from LITTERA, but rather that scripted speech is an interpretation by a given narrator and not "organic" per se.

Because of LITTERA's pedagogical aims, as well as pragmatic issues throughout its compilation, the issue of representativeness is not as important as the question underlying anything designed for pedagogical purposes; that is, *what useful information can this tool provide students?*, the answers to which are explored in chapter 3, Pedagogical Applications.

2.7 LITERATURE, LANGUAGE LEARNING AND LITTERA

This chapter will conclude with a brief discussion regarding some of the research into literature and language learning given that it is this text type that comprises the LITTERA corpus.

McKay (1982) outlines three common arguments against the use of literature in language learning. These are: 1) that its "unique use of language" does not provide students with adequate models of grammar, which is the main goal of language teaching; 2) that it is not useful for students' long-term goals, whether academically or professionally; 3) because literature "often reflects a particular cultural perspective", it is therefore likely too be too difficult for students from another culture and language to understand and appreciate. While she addresses all three in her article, it is the first that is most pertinent to the present work. In her response to this criticism, she cites Widdowson's (1978) two levels of linguistic knowledge, *usage* and *use*. *Usage* refers to one's understanding and awareness of the linguistic rules and structure while *use* refers to knowing how to effectively apply the rules of *usage* in communicative situations, akin to Saussure's famous distinction between *langue* and *parole* or Chomsky's *competence* and *performance*. According to McKay, knowledge of *usage* can be gained through salient grammatical features within the texts while vocabulary is expanded through attention to the surface form of words. As for *use*, she points out that "literature presents language in discourse" and "language that illustrates a particular register or dialect is embedded within a social context, and thus there is a basis for determining why a particular form is used". In other words, literary texts provide examples of how language is being used in a given context and raise questions about intentionality and purpose to the reader. Not only is there interaction between characters in the text through dialogue, there is also the interaction of the reader with the text itself, which puts the question of *use* at the forefront, continuously opening itself up to examination.

McKay is not alone in her defense of literature in language learning. There is now a substantial body of research (Tannen, 1982; Brumfit & Carter, 1986; Cook, 1994; Kasper, 2000; Hall, 2005; Johns et al., 2008; Hernández, 2011) which supports the claim that the language found in literature is neither irrelevant nor beyond students' grasp. Much of this research challenges the notion that there is such a clear-cut thing as *literary language*. Therefore, to better understand what literary texts can contribute to language learning, it is necessary to get at the underlying concept of *literary language*.

The idea of *literary language* as a completely separate and unique form of discourse has been challenged repeatedly. Hall (2005:10) posits that "'literariness' is a matter of degree rather than kind" and goes on to explain that there is a "surprising degree of literariness of the ordinary, and the equally pervasive ordinariness of the literary, particularly in the modern period" (11) and adds, "nor…is there any indication that 'literary' language as opposed to 'ordinary' language is an empirically valid distinction" (23). This is because it is not language per se that is particularly unique in literary works, but rather what is being done with language. Techniques such as deviation, foregrounding and parallelism are what make ordinary language literary. They draw the readers attention to the way language is being used and why that is significant. While there is certainly a much higher degree of planning that goes into literature than spontaneous speech, both forms of discourse share certain strategies and techniques. One of those is metaphor. Many researchers (Tannen, 1982; Brumfit & Carter, 1986; Lakoff, 1993; Wieser, 2005) have pointed out how pervasive metaphors are in our everyday language use. Tannen also points out other literary devices that comprise both literary language and spontaneous conversation, such as "the repetition of sounds and words, syntactic structures and rhythm". She goes on to explain:

> "The written short story…takes advantage of the written medium to achieve integration, to create maximum effect with fewest words; but it depends for its impact, like face-to-face conversation, on a sense of involvement between the writer and the audience or characters in the narrative. It is for this reason that literary discourse (short stories, poems, novels), rather than being most different from ordinary conversation, is in fact most similar to it: those features which are thought quintessentially literary (repetition of sounds and words, syntactic parallelism, rhythm) are all basic to ordinary spontaneous conversation"

Taking this into account, then, it makes sense that literary texts would be of interest to language learners. Literary works expose the learner to a wide range of styles, voices and registers (Biber, 1988; Hall, 2005). As Tan (2013) writes, "[t]he parasitic relationship between literary texts and other kinds of discourse means that it is possible to discuss general linguistic and pragmatic phenomena through examining literary texts". In addition, literary texts are often rich in description, which provides the learner with an array of collocations that can then be analyzed through a corpus-based approach. At the moment there is no way to explore collocations directly from the LITTERA interface, although a search for, say, *hair* will provide results indicating related nouns and common descriptors which may prove useful for students, such as a *lock of hair*, *braided hair*, *brushed hair*, *strand of hair*, *soft hair*, *straight hair*, etc. This may serve as a springboard for follow-up searches for related terms from the results, such as *braided*, which yields seven results from five different texts, all referring to *hair*. Many such collocations for body parts, physical spaces, movements and so on can be searched in the corpus as a whole or within an individual text.

Then there is the fact that literary texts frequently contain representations of spontaneous speech through dialogue. From a phonological standpoint, this may lead the narrator to display certain features of conversational language that may not come through when reading prose. Authors also attempt to use orthography to represent informal speech, especially in more modern works as will be seen in section 3.2.1.7 below. This technique provides the learner with insights regarding certain characteristics of conversational English, such as the dropping of *g* in many gerunds, e.g. *walkin'* or *talkin'*, which represents the change from /ŋ/ to /n/ in word-final position. Furthermore, a corpus-based analysis of dialogue makes it easier for the user to locate examples of direct and indirect speech[27], a common topic in ESL classrooms, along with other features of spoken language, such as discourse markers, hedging, fillers, tag questions, etc.

However, it is important not to equate literary dialogue with that of real conversations, despite much overlapping and many shared features. As Tan points out, "the tedious repetitions, false starts and reformulations that characterize spontaneous conversations will not be fully represented in literary conversations". Turning spontaneous speech into written language will usually mean that certain features of spoken language, such as gestures and voice quality, will be lost (Amador-Moreno, 2010), along with the aforementioned features from Tan. Nevertheless, Amador-Moreno adds that the dialogue represented in fictional works is "rooted in ordinary discourse and everyday situations", rendering it useful for language learners.

In terms of difficulty, Hall (2005) notes that it is vocabulary that has been shown to be a more determining factor in literary text difficulty than syntax. While this issue can be solved with a dictionary, being able to search within the text via a parallel corpus provides additional advantages. Besides the obvious benefit of being able to examine the translation of a word within the same context, students can see how often a word is used, which would indicate how important it will be for comprehension within that text. Word frequency can also be compared to the corpus as a whole. If a word is infrequent in one text but frequent in the corpus overall, it still may be an item the student will want to focus on.

Lastly, a corpus comprised of literary texts such as LITTERA may prove to be a useful resource for what has come to be known as *corpus stylistics*. Corpus stylistics pairs the methodologies and techniques of corpus linguistics with the study of stylistics, which McIntyre & Walker (2019:16) define as "the linguistic study of style in language and how this is influenced

[27] Direct speech refers to direct quotations such as *Did you take off your shoes?*, while indirect speech refers to how another person narrates the same utterance, e.g. *John asked Pam if she had taken off her shoes*.

by particular non-linguistic variables" including the author's biography, genre, historical context, geographical location and so on. They add that, more broadly, stylistics looks at "how the linguistic choices evident in a text contribute to the overall meanings and effects of that text". Both corpus linguistics and stylistics take an empirical approach to language analysis. Stylistics incorporates theories from a number of areas within linguistics, such as phonetics, semantics, syntax, cognitive linguistics, historical linguistics, pragmatics and sociolinguistics (Wynne, 2006), while corpus linguistics provides the tools with which an empirical examination of a given text can be carried out. Corpus linguistics therefore provides the stylistician with additional resources to perform a stylistic analysis, allowing him or her to prove or disprove an originally qualitative analysis of a text. McIntyre & Walker go on to provide examples of style that can be analyzed through corpora, such as semantic preference (via collocations), the use of exclamation marks signaling emotions, the use of irony (cued by specific words and phrases), politeness, conversational structure, visual features, syntax and lexis. Furthermore, in her study on metaphor in the SCOTS corpus, Anderson (2013) shows how metaphors can be examined in a variety of texts using corpus techniques. Such a cursory overview of corpus stylistics barely scratches the surface of what is becoming a fascinating and complex field of study. For more on corpus stylistics see Wynne, 2006; Amador-Moreno, 2010; Shepherd & Sardinha, 2013; McIntyre, 2015; McIntyre & Walker, 2019.

The current subchapter has attempted to present a brief look at literature in language learning and how it pertains to the LITTERA corpus and corpus linguistics more generally (e.g. corpus stylistics). The notion of *literary language* was discussed along with a look at some of the advantages of literary texts in language learning. Hernández (2011) provides an adequate summary of the many benefits of literary texts in the language classroom:

> "It is generally agreed that literary texts seem to be an ideal tool both for developing literary comprehension and sensibility and also to enhance the communicative skills of the language: literary texts supply examples of authentic language, provide lots of opportunities for the expression of ideas, opinions, and beliefs and are a springboard for any writing activity. Furthermore, literature helps enhance the psycholinguistic aspect of language learning as it focuses on form and discourse processing skills and

> improves vocabulary expansion and reading skills. The text seems to be the perfect vehicle to investigate the stylistic features of an author and the characteristics of a period. However, texts can also be explored at different levels: they can present information about culture and society and may be analysed with different purposes."

It should be clear, then, that a corpus of literary works will certainly be conducive to language learning in its own right, especially with texts that students will likely become familiar with throughout their English philology coursework or that they may have already read on their own.

This chapter has looked at the conception, development and composition of the LITTERA corpus, along with other relevant aspects such as related corpora, a general profile of the target user and, lastly, a brief overview of literature in language learning. This audio-textual English-Spanish parallel literary corpus is made up of literary texts based on the curricula for the English philology program at the University of Santiago de Compostela so as to be pedagogically relevant to the students of the program. Furthermore, audio from audiobooks have been added with the aim of addressing phonology through the lens of Data-Driven Learning, which allows for users to explore a variety phonological features of English directly in the corpus data, examples of which will be presented in the next chapter.

3. PEDAGOGICAL APPLICATIONS

Now that a detailed description of LITTERA’s creation, design and features has been provided, it is possible to begin an analysis of how the corpus can be used in language pedagogy by Spanish learners of English, particularly in the study of English phonology. This chapter will examine the corpus data for what it can tell students about speech production and comprehension.

Firstly, a series of training exercises will be outlined in detail in section 3.1. These exercises are meant to familiarize users with LITTERA’s content and search features. Section 3.2 will then analyze the phonetic data in the corpus and discuss what insights students can gain, laying out *how* the corpus data can be exploited for language learning. In doing so, both broad suggestions for classroom activities and examples of specific searches will be presented that will allow for a well-rounded examination of the different phonological phenomena in connected English speech. It is the author’s hope that this analysis of LITTERA’s pedagogical potential inspires future empirical research to be carried out on the effectiveness of said data in language learning, whether through the LITTERA corpus or any other speech corpus.

3.1 Training Exercises with LITTERA

Before proceeding to the actual corpus work, it is worth taking time to describe a series of ‘training’ exercises meant for users to familiarize themselves with the corpus and become acclimated to the different search options available. The first set of training exercises will be through the CLUVI interface, while the second set will be through the SensoGal interface. Much of the work below is inspired by Frankenberg-Garcia’s (2012) work on training teachers in basic corpus skills and the DDL approach, as discussed previously in section 1.2.4.

The training exercises described below are taken from the video tutorials available on the respective homepages for both versions of the LITTERA corpus. Each set of videos is divided into three parts consisting of a general overview along with further practice with the various search options. The scripts for the tutorials can be found in Appendices B & C. All videos are narrated in English with English and Spanish subtitles available. Users are asked a question, then expected to pause the video to find the answer, which is then given and explained.

3.1.1 Set 1: CLUVI

The LITTERA tutorial series for CLUVI is divided into the following three sections: Corpus Overview (3.1.1.1), Simple Searches (3.1.1.2), and Complex Searches with Regular Expressions (3.1.1.3). The Corpus Overview questions are aimed at familiarizing the user with the basic features of the corpus, such as word count, languages, numbers of translation units, the literary works and their reference codes, etc. The Simple Searches tutorial introduces the user to the basic search functions while the Complex Searches with Regular Expressions tutorial takes it a step further and provides exercises meant to practice some of the most common regular expressions available in the corpus. Hopefully through this kind of 'guided tinkering', the user is able to adequately exploit the corpus in a relatively short period of time.

3.1.1.1 Corpus Overview

Users can begin at the LITTERA homepage[28] where they will find all the necessary information to answer the following exercise questions. The questions in this first part of the tutorial series may perhaps seem oddly simple, but the idea is to make the user aware of the corpus's size and composition so that this is kept in mind when carrying out searches and analyzing the data.

The first question is meant to point out the total number of words and how the words are distributed between languages:

1. How many total words are in the corpus? How many per language?

The user is pointed to the description on the homepage, which shows that there are nearly two million words in the corpus: 1,968,676 to be exact—983,618 in English and 985,058 words in

[28] http://sli.uvigo.gal/CLUVI/index.php?corpus=24&tipo=16&lang=en

Spanish. A brief explanation is then provided on corpora more generally, such as factors that often determine a corpus's size and content, as well as the origin of most of the texts included in the corpus (English philology syllabi).

The second question introduces the concept of translation units:

> 2. According to LITTERA's homepage, how many translation units are in the corpus?

This answer, 63,508, can be found next to the total number of words. The definition of a translation unit (a bilingual pairing that contains the same information in different languages) is also provided.

In order to make the user aware of how texts are categorized, the third question asks:

> 3. What is the three-letter reference code for Virginia Woolf's *The Mark on the Wall*? And for Ernest Hemingway's *The Old Man and the Sea*?

This exercise makes users aware of the three-letter code for each text as it is this code that will allow the user to identify which text a given TU comes from in the search results.

The remaining questions in the overview are intended to make the user aware of proportionality and distribution in the corpus and how the number of translation units is not strictly contingent on the number of words. For example, the user may see that *Sense & Sensibility* has 4,681 translation units and therefore make the assumption that it makes up less of the corpus than, say, *Lord of the Flies*, which has 5,555 translation units. Yet, *Sense & Sensibility* has double the amount of words in *Lord of the Flies*, 118,503 to 59,128 respectively. The same goes for the text with the most TU's, *The Hunger Games* (6,628 TU's and 99,762 words). The user is then asked to consider why this may be.

> 4. How many English words are in *The Lord of the Flies*? How many translation units?
> 5. Which book contains the most words in English? The least?
> 6. For which book are there the most translation units? The least?

7. Why might the book with the most words not have the most translation units?

The differences between the text with the most words (*Sense & Sensibility*) and the text with the most TU's (*The Hunger Games*) tells us that each translation unit in *Sense & Sensibility* will, on average, contain more words than those from *The Hunger Games* or *Lord of the Flies.*

3.1.1.2 Simple Searches

After the size and composition of the corpus have been examined, the user can move on to part two, which focuses on forming simple search queries.

A good way to begin is to search for a single word one language at a time. First, users are told to turn their attention to the search menu and it is noted in the video that, by default, the *Lexical Equivalents* option is automatically ticked, the *Results* option is set to 20 and the *Spoken English Variety* option is set to *All*. The user is then asked to carry out a simple search:

1. Without adjusting any of the default settings, how many total results are there for *play*?

After explaining where the answer, 155, can be located on the page, other aspects of the search results display are pointed out, such as the Galnet dictionary entry (as a result of the Lexical Equivalents option being ticked by default), the highlighting of equivalents in green in the translation, the three-letter text identification code (and how it can be used for quick access to the bibliography), the TU's index number within the text, the play and rewind buttons, and finally the ability to change the display format (from horizontal to vertical).

The user is then asked to carry out the same search, but now un-ticking the Lexical Equivalents option.

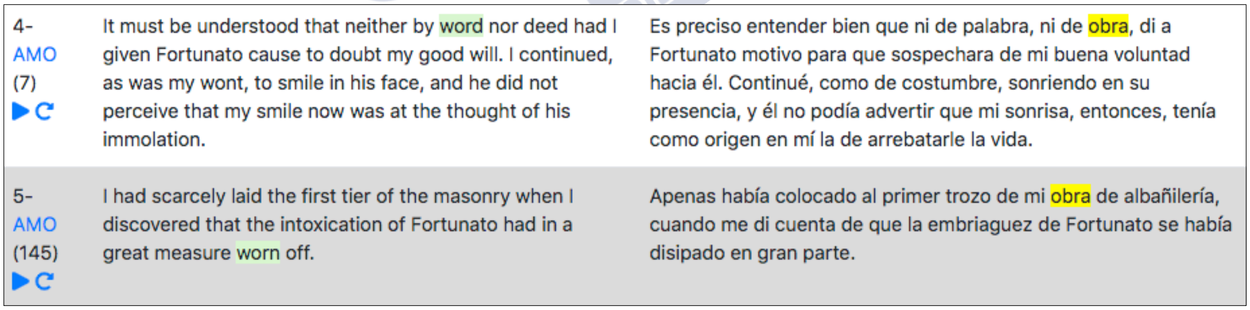
2. Uncheck the *Lexical Equivalents* option. How many total results are there for *play* now? Why is this number different than the previous search? (Feel free to adjust the Wider Context option or the number of results displayed. However, do not adjust the Spoken English Variety option as it will omit all results in which the spoken variety is not the selected one.)

The search now yields 423 examples because, in this case, all the examples containing the string *p-l-a-y* are included in the results, such as *played*, *display*, *playing*, etc. Furthermore, it is noted that the Galnet dictionary entries are no longer present in the header with the *Lexical Equivalents* option turned off. Nor are any of the lexical equivalents highlighted in green.

The previous two searches are then repeated in Spanish with the word *obra*. The first is with the Lexical Equivalents ticked:

> 3. Leaving the English search box empty but checking the *Lexical Equivalents* option, how many results are there if we search for *obra* in Spanish?

The search yields 53 results, although now only some of the lexical equivalents are highlighted in green. It is explained that this is due to the fact that no words from the Galnet dictionary entry at the top were found in the English text. It is also pointed out that in the fourth and fifth results, the lexical equivalent has been incorrectly assigned. In the fourth result, *word* is highlighted in green when it should be *deed*. In the fifth result, *worn* is incorrectly highlighted when, in reality, there is no direct equivalent in the translation, as seen in Figure 25 below.



4- AMO (7) ▶ ↻	It must be understood that neither by word nor deed had I given Fortunato cause to doubt my good will. I continued, as was my wont, to smile in his face, and he did not perceive that my smile now was at the thought of his immolation.	Es preciso entender bien que ni de palabra, ni de obra, di a Fortunato motivo para que sospechara de mi buena voluntad hacia él. Continué, como de costumbre, sonriendo en su presencia, y él no podía advertir que mi sonrisa, entonces, tenía como origen en mí la de arrebatarle la vida.
5- AMO (145) ▶ ↻	I had scarcely laid the first tier of the masonry when I discovered that the intoxication of Fortunato had in a great measure worn off.	Apenas había colocado al primer trozo de mi obra de albañilería, cuando me di cuenta de que la embriaguez de Fortunato se había disipado en gran parte.

Figure 25. Search results showing incorrectly assigned lexical equivalents.

The same search is then repeated in the next exercise, but without the Lexical Equivalents option ticked.

> 4. Repeat the same search as in exercise 3, but now without the *Lexical Equivalents* option checked. How many results are there?

The user will see the same thing happen as in question two. The search treats *obra* like a string of characters, *o-b-r-a*, rather than a lexical item, thus returning such results as *sobra*, *obras* and *cobrar*. The user is then asked to follow up these single language searches with a search in both languages using *play* and *obra*.

> 5. Now search for *obra* in Spanish and *play* in English simultaneously. How many results are there? (Note: when both languages are searched simultaneously, the *Lexical Equivalents* option no longer applies, whether it is checked or not.)

This search yields 12 results and both terms are highlighted in yellow. The Lexical Equivalents option has no effect on the results and no dictionary entry appears at the top of the page due to the fact that this option only functions when searching by lemma in one language at a time.

The final exercise in part two is meant to show the user that it is possible to search within a single text and compare those results to those from the corpus as a whole.

> 6. Follow the menu options so that you are only searching in George Orwell's *1984* (Full-text search → Audiobooks → English-Spanish → George Orwell – *1984*). Leaving the *Lexical Equivalents* option checked, how many times does the word *party* appear in *1984*? How many times does it appear in the entire corpus?

Party appears 259 times in *1984* and 425 times in the corpus as a whole (including those from *1984*). This means that around half of the results for *party* in the entire corpus are from a single text, making the user aware of proportionality in the data.

3.1.1.3 Complex Searches with Regular Expressions

Part three deals exclusively with regular expressions. As seen in Figure 3 in the previous chapter, the *Help* page provides an explanation of all the regular expressions available to the user. But rather than learning every single one, it is arguably more advantageous to first introduce the user to some of the more practical regular expressions, which can be found in Table 2. These were selected on the basis of being, in the author's opinion, the most useful for beginners[29]:

Table 2. Regular expressions used in the tutorial on Complex Searches within CLUVI.

Regular expression	Function
\b	Marks a word boundary
[ ]	Searches any of the contained characters
|	"or" (e.g. *shoe|sock* searches for all instances of *s-h-o-e* OR *s-o-c-k*)
-	Range: from one character in a sequence to another (e.g. *s-v* refers to *s*, *t*, *u*, and *v*.)
?	Show one or zero instances of preceding character (e.g. *shoes?* returns the singular and plural forms of *shoe*)
\	Make character literal (i.e. when characters that function as regular expressions are meant as their original character, such as the question mark. Therefore, *\?* will find all question marks in the corpus.)
\w	Represents any character that can form part of a word
*	Zero or more of the previous character

Mastering a small handful of regular expressions will provide users with the skills and confidence to eventually explore the others. It should be noted that by choosing these regular expressions specifically the author is not implying that the others are less valuable or will not be used. These have been chosen due to the fact that they do not require much elaboration and will likely prove useful for ordinary searches and will be relevant to the pedagogical applications described below in section 3.2.

The exercises here are meant to achieve two objectives: 1) familiarize users with the different regular expressions listed in Table 2 by employing them in the exercises and making inferences based on observation, and 2) have users then apply their new knowledge to practical exercises that may require the use of multiple regular expressions in a single search.

The first search is meant to introduce the idea of a word boundary. Note that there is no reference to the Lexical Equivalents option in these exercises. This is because no corresponding term will be located in Galnet as the program will not recognize anything that is not a lemma. Therefore, whether it is ticked or not is immaterial.

[29] In retrospect, the caret ^ , which when used within brackets signifies exclusion, should have also been added to this tutorial as it will be used regularly in the analysis of suprasegmentals below.

1. Place *\b* before the word *play* (without a space in between *\b* and the word) and carry out a search. Judging by the results, what does *\b* do?

Along with an explanation of what the boundary does and how only words beginning with *play* such as *played* or *playing* will appear in the results, it is also explained that another way to create a boundary is to place a space before the word. However, it is made clear that this will exclude any instances of *play* at the beginning of a translation unit or a quote. Therefore, a boundary includes all instances of *play-*, regardless of what precedes it.

The second question asks the user to infer what brackets do based on a specific search.

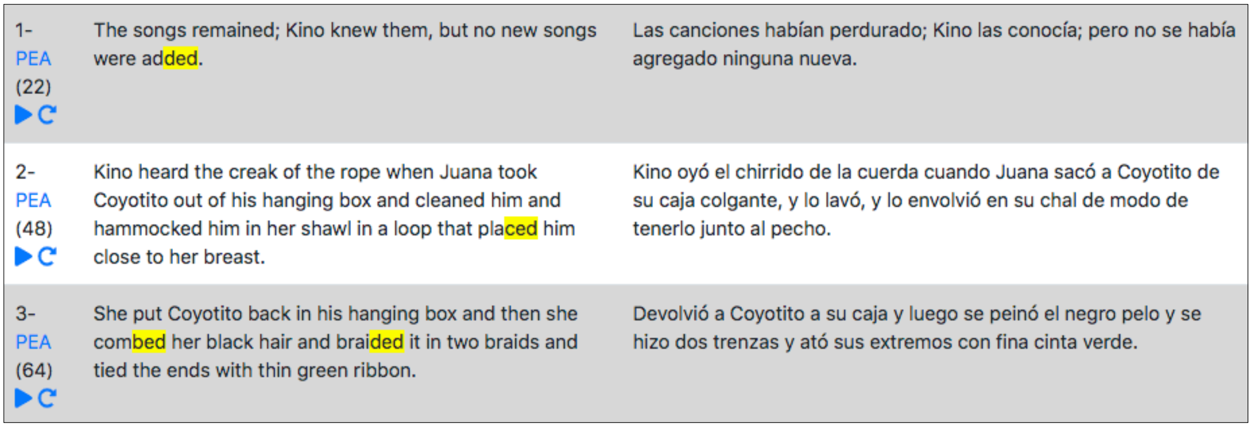
2. What do brackets ([ ]) do judging by the results from the query *ma[kd]e sure*?

Because brackets search for zero or one instance of any character contained within, this search will return all instances of *make sure* and *made sure*. It is then explained that adding *-r-* to the brackets in this search would not have any effect on the results as *mare sure* is unsurprisingly not found in the corpus data. This is a useful regular expression in order to include both the present and past tense of irregular verbs in English, as seen here, or for stem-changing verbs in Spanish, as mentioned above in section 2.5.1 with the verb *pedir*.

The third question shows the user how to mark a range:

3. What does a dash (–) do judging by the results from the query *[b-d]ed\b*?

In this case the dash includes results for words ending in *–bed*, *–ced* and *–ded* as seen in the search results in Figure 26 below.



1- PEA (22) ▶ C	The songs remained; Kino knew them, but no new songs were added.	Las canciones habían perdurado; Kino las conocía; pero no se había agregado ninguna nueva.
2- PEA (48) ▶ C	Kino heard the creak of the rope when Juana took Coyotito out of his hanging box and cleaned him and hammocked him in her shawl in a loop that placed him close to her breast.	Kino oyó el chirrido de la cuerda cuando Juana sacó a Coyotito de su caja colgante, y lo lavó, y lo envolvió en su chal de modo de tenerlo junto al pecho.
3- PEA (64) ▶ C	She put Coyotito back in his hanging box and then she combed her black hair and braided it in two braids and tied the ends with thin green ribbon.	Devolvió a Coyotito a su caja y luego se peinó el negro pelo y se hizo dos trenzas y ató sus extremos con fina cinta verde.

Figure 26. Search results for *[b-d]ed/b*.

The fourth question is the first in part three in which the user is asked to elaborate his or her own search query based on information from earlier in the video.

4. How can we elaborate the previous search, *[b-d]ed\b*, with dashes to return all instances where a consonant appears before the past tense *-ed* morpheme (excluding *-w-* and *-y-*)?

Such a query can be formulated as *[b-df-hj-np-tvxz]ed\b*. *-W-* and *-y-* are excluded due to the fact that in word-final position, they would lead to a vowel preceding the past tense morpheme, as in *clawed* and *played*.

In the fifth question, the vertical pipe is introduced:

5. What does a vertical pipe (|) do judging by the results from the query *make sure|made sure*?

The vertical pipe is an "or" statement, which means that this search query translates to "find all instances of *make sure* **or** *made sure*". The query returns the same number of results as the search using brackets, which shows users that the same results can be achieved with different regular expressions. It is often the case that there are numerous ways to yield a certain set of results.

The sixth question introduces the question mark as a regular expression.

6. What does the question mark (?) do judging by the results from the query *\bci?eg* in the Spanish search box?

The question mark returns zero or one instance of the preceding character, which is why *ciego* (one instance of -*i*-) and *cegarlo* (zero instances of -*i*-) both appear in the results.

The seventh question, much like the fourth, attempts to get students to formulate the query for themselves. In this case, in order to optimize the search results, they will have to use four of the aforementioned regular expressions in order to deal with a stem-changing verb.

7. How can we return all possible forms of the stem-changing verb *negar* (*negando*, *niego*, *negó*, *negué*, *negaron*, *niegue*, etc.) while minimizing any unwanted results? (Note: this may take a few attempts to find the most optimal search.)

The recommended query to return all possible forms of *negar* is ***\bni?eg[aóu]|\bniego\b***. The least intuitive part of this query is separating *niego* with a vertical pipe (i.e. an “or” statement). This is done to prevent the extraneous results caused by *negociar* and its derived forms such as *negociador, negocios*, *negociado*, etc, from overwhelming the search results. Nevertheless, extraneous results will still slip through the cracks, even with the recommended search, particularly *negativa* and *negativas*. This can be taken care of, although it will require the use of another regular expression that is not dealt with in the tutorial; that is, the caret symbol “ ^ ”, which when used in brackets functions as a *not* statement. The search would then be ***\bni?eg[aóu][^t]|\bniego\b***. Because this regular expression is not described in the video, this alternative is not included as the extraneous results are minimal and do not hinder the corpus search.

The eighth question incorporates both languages into the search query:

8. How many results are there if we want to return all forms of *negar* when they correspond to the English verb *shake* (past form: *shook*)?

While the same query from the previous exercise is recommended for the Spanish query, the English query can be expressed as ***shook|shak***. The -*e*- in *shake* is omitted so as to not exclude the gerund form of the verb, *shaking*.

The ninth question focuses on the backslash, which takes a character that on its own functions as regular expression and makes it literal.

9. What does the backward slash (\ ) do judging by the results from the queries *\?* or *\]*?

This query will return all instances of question marks and brackets within the text. The same goes for other characters that otherwise function as regular expressions, such as periods (.), asterisks (*) and the plus sign (+). The video then goes on to explain that the number of translation units in the results for *\?* does not equate to the number of questions in the corpus as there are likely a number of TU's with more than one question mark. Lastly, it is noted that with certain letters a backslash has the opposite effect, as seen above with *\b*, taking a letter and turning it into a regular expression. This also occurs in the next question:

10. What does *\w* do judging by the results from the queries *\b\wat\b* and *\b\w\wat\b*?

When a backslash is placed in front of -*w*-, -*w*- becomes a regular expression representing any character that can form part of a word. That is why results such as *eat*, *mat*, *pat* and *fat* are returned with the first query, and *that*, *goat* and *beat* are returned with the second. The only thing these words have in common is the number of letters designated by the search query and the fact that they all end with –*at* as specified.

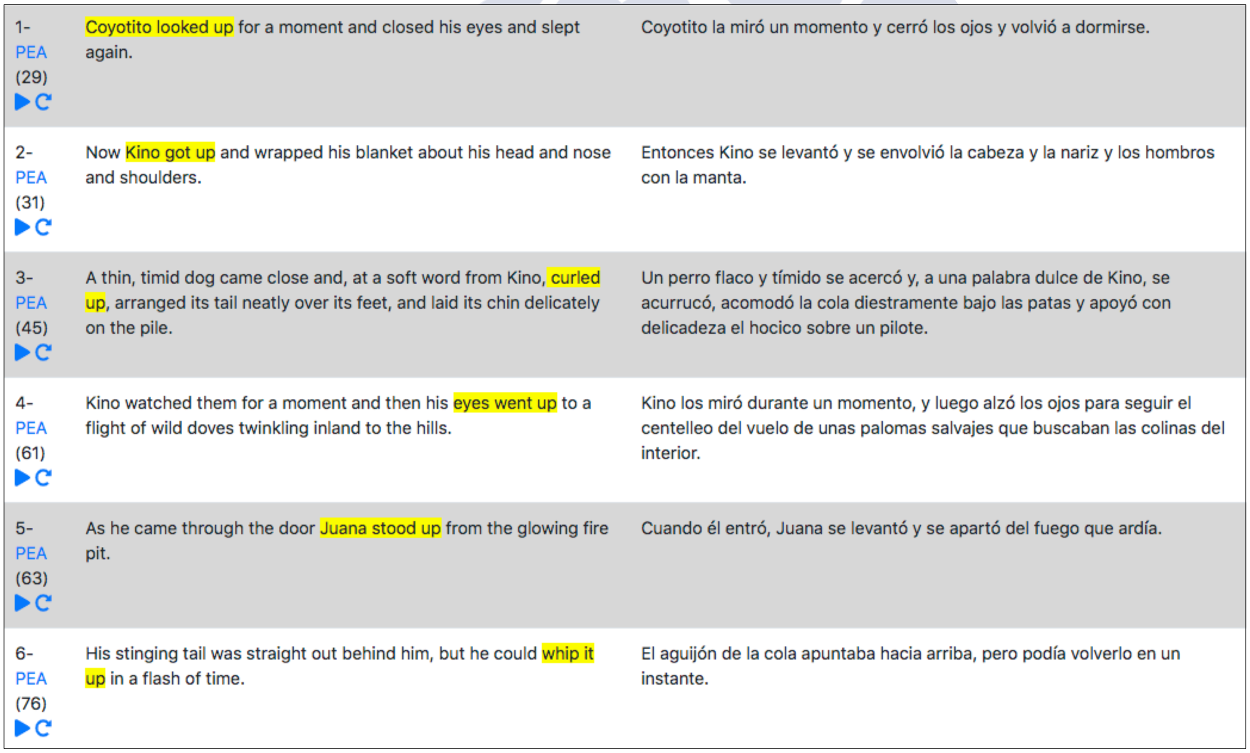
The last regular expression presented in the video tutorial is the asterisk, which represents zero or more of the previous character.

11. What does adding an asterisk after *\w* do in the query \b\w*at\b?

This search returns any word that ends in *–at* regardless of the number of letters. Another way to find words ending in *–at* is to begin the query with -*a*- and place a boundary after -*t*-, as in *at\b*. However, the asterisk is useful in other ways, as seen in the final question of the video:

> 12. How can we search for all sequences of three words in which the third word is always *up* (e.g. *give it up* or *he looked up*)?

The simplest way to indicate a separate word is to use \w* separated by a space. Therefore, the recommended query is **\w*_\w*_up\b** (henceforth, underscores will be used to signal spaces within search queries). This is a useful way to search for multiple word strings in which one or more words is to be left unspecified, such as with English phrasal verbs. The results from this search show that searching by particle is an effective way to locate phrasal verbs in the data.



1- PEA (29) ▶ ↻	Coyotito looked up for a moment and closed his eyes and slept again.	Coyotito la miró un momento y cerró los ojos y volvió a dormirse.
2- PEA (31) ▶ ↻	Now Kino got up and wrapped his blanket about his head and nose and shoulders.	Entonces Kino se levantó y se envolvió la cabeza y la nariz y los hombros con la manta.
3- PEA (45) ▶ ↻	A thin, timid dog came close and, at a soft word from Kino, curled up, arranged its tail neatly over its feet, and laid its chin delicately on the pile.	Un perro flaco y tímido se acercó y, a una palabra dulce de Kino, se acurrucó, acomodó la cola diestramente bajo las patas y apoyó con delicadeza el hocico sobre un pilote.
4- PEA (61) ▶ ↻	Kino watched them for a moment and then his eyes went up to a flight of wild doves twinkling inland to the hills.	Kino los miró durante un momento, y luego alzó los ojos para seguir el centelleo del vuelo de unas palomas salvajes que buscaban las colinas del interior.
5- PEA (63) ▶ ↻	As he came through the door Juana stood up from the glowing fire pit.	Cuando él entró, Juana se levantó y se apartó del fuego que ardía.
6- PEA (76) ▶ ↻	His stinging tail was straight out behind him, but he could whip it up in a flash of time.	El aguijón de la cola apuntaba hacia arriba, pero podía volverlo en un instante.

Figure 27. Results for *\w*_\w*_up\b*. Note the phrasal verbs.

This concludes the three-part tutorial series for the CLUVI interface. Part one provided an overview of LITTERA's size and composition while introducing some basic corpus concepts such as translation units and proportionality. Part two presented the basic features of both the search menu and the results page through a series of simple searches. Part three introduced the idea of regular expressions to form complex search queries, allowing the user more flexibility to explore the corpus data.

3.1.2 Set 2: SensoGal

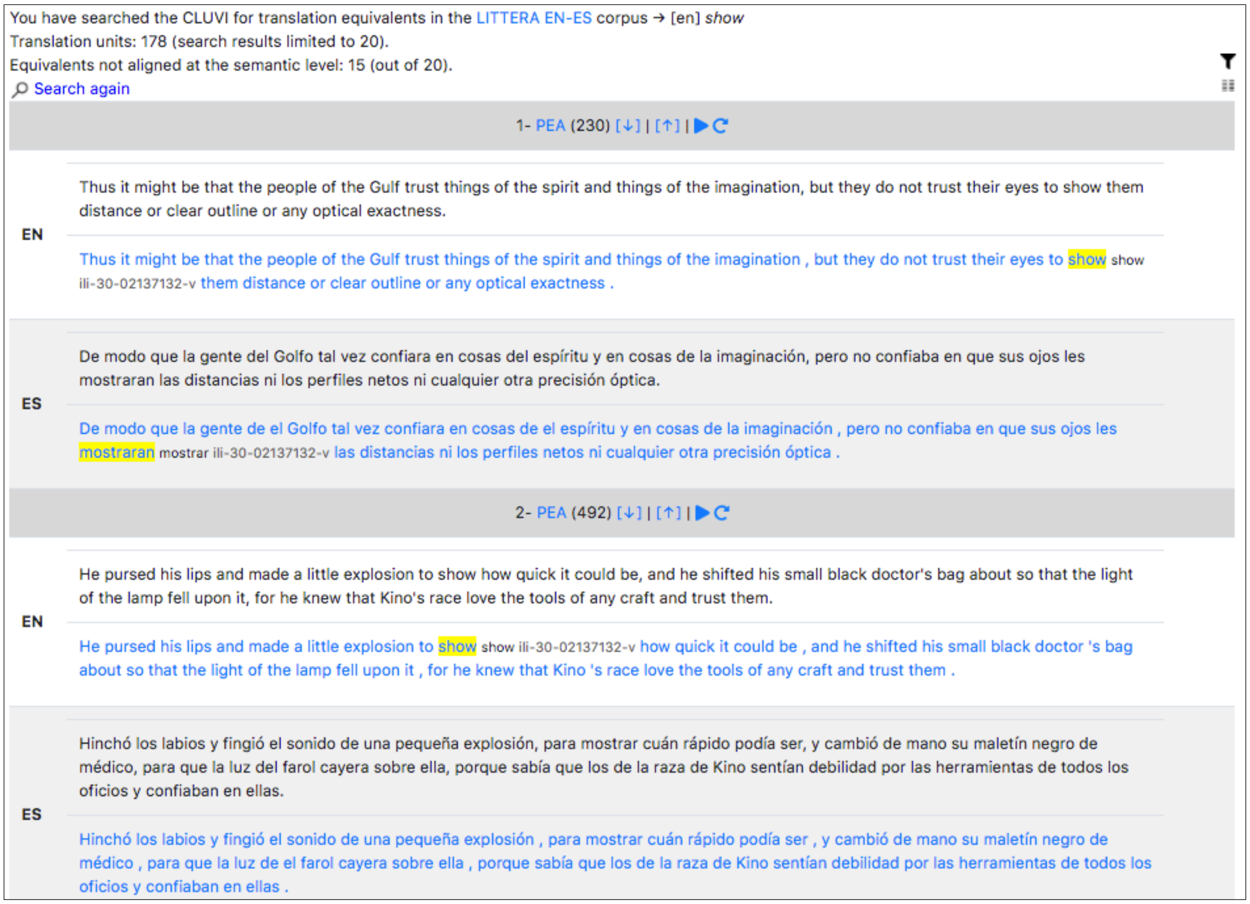
Exploring LITTERA through SensoGal allows for certain options that are unavailable when accessing the corpus through CLUVI due to the fact that the corpora within SensoGal, including LITTERA, have been semantically annotated. The SensoGal tutorials are also divided into three parts: 1) General Overview, 2) Search by Form and 3) Search by Concept. Because the General Overview tutorial is practically identical to that previously described in the CLUVI set, we will begin with the second part, Search by Form.

3.1.2.1 Search by Form

Similar to the simple search exercises for LITTERA via CLUVI, users are asked to turn their attention to the different options in the search menu, now divided into two sections, *Search by form* and *Search by concept*. While many of the same options are available as in the CLUVI menu when searching by form, users can now also specify the part of speech. There are two ways to search by form: by word or by lemma. Therefore, the user is first asked to search by word by unchecking the *Search Lemma* option:

1. Uncheck the *Search Lemma* box but leave all the other default settings as they are. Type the word *show* into the English search box and carry out a search. How many total translation units are found (not necessarily shown) and how many equivalents were not aligned at the semantic level (see the information located above the first result near the top of the page)?

The search yields 178 translation units and shows that 15 equivalents were not aligned at the semantic level. The objective here is to draw the users attention to the basic features of a search query and the results page, and the fact that searching by word will only yield the word entered in the query box and no other forms. The tutorial goes on to explain what a translation unit is now that it can be visualized on the screen, and also what it means that semantic equivalents could not be aligned. Because the idea of semantic equivalents is not immediately apparent, an explanation is given on how to make sense of the results, since, unlike those for the CLUVI interface, now there are two examples of the text in each translation unit: one for the original text, in black, and another showing the semantic annotations, in blue, as seen in Figure 28 below.



You have searched the CLUVI for translation equivalents in the LITTERA EN-ES corpus → [en] *show*
Translation units: 178 (search results limited to 20).
Equivalents not aligned at the semantic level: 15 (out of 20).
Search again

1- PEA (230) [↓] | [↑] | ▶ ↻

EN

Thus it might be that the people of the Gulf trust things of the spirit and things of the imagination, but they do not trust their eyes to show them distance or clear outline or any optical exactness.

Thus it might be that the people of the Gulf trust things of the spirit and things of the imagination , but they do not trust their eyes to show show ili-30-02137132-v them distance or clear outline or any optical exactness .

ES

De modo que la gente del Golfo tal vez confiara en cosas del espíritu y en cosas de la imaginación, pero no confiaba en que sus ojos les mostraran las distancias ni los perfiles netos ni cualquier otra precisión óptica.

De modo que la gente de el Golfo tal vez confiara en cosas de el espíritu y en cosas de la imaginación , pero no confiaba en que sus ojos les mostraran mostrar ili-30-02137132-v las distancias ni los perfiles netos ni cualquier otra precisión óptica .

2- PEA (492) [↓] | [↑] | ▶ ↻

EN

He pursed his lips and made a little explosion to show how quick it could be, and he shifted his small black doctor's bag about so that the light of the lamp fell upon it, for he knew that Kino's race love the tools of any craft and trust them.

He pursed his lips and made a little explosion to show show ili-30-02137132-v how quick it could be , and he shifted his small black doctor 's bag about so that the light of the lamp fell upon it , for he knew that Kino 's race love the tools of any craft and trust them .

ES

Hinchó los labios y fingió el sonido de una pequeña explosión, para mostrar cuán rápido podía ser, y cambió de mano su maletín negro de médico, para que la luz del farol cayera sobre ella, porque sabía que los de la raza de Kino sentían debilidad por las herramientas de todos los oficios y confiaban en ellas.

Hinchó los labios y fingió el sonido de una pequeña explosión , para mostrar cuán rápido podía ser , y cambió de mano su maletín negro de médico , para que la luz de el farol cayera sobre ella , porque sabía que los de la raza de Kino sentían debilidad por las herramientas de todos los oficios y confiaban en ellas .

Figure 28. Search results for *show* (searched by form) in part two of the SensoGal video tutorials.

It is the text in blue that must be further explored. It is explained that each annotated lexical word is accompanied by an Interlinguistic Index (ILI) code. By clicking on this code, the user is redirected to Galnet, which is also explained in slightly more detail. The concept corresponding to *show* in the first TU is *make visible or noticeable*. Other lexical equivalents in the other languages, particularly Spanish, are then pointed out. Because *show* and *mostraran* correspond to the same concept in the first TU, they share the same ILI code. Yet, in the second TU, there is no lexical equivalent highlighted in Spanish. This is what *not aligned at the semantic level* is referring to. Therefore, to hide all instances of non-matches, the user is told to click on the funnel icon in the upper right.

The same general features of the search results are then explained, as in 3.1.1.1 above, pointing out the results number, the three letter reference code for each text, the TU index number, etc. Another difference between CLUVI and SensoGal is the set of arrows above each TU, which show and hide all the ILI codes in the TU.

Because the previous search was carried out by word, the user is now told to check the *Search Lemma* option and repeat the same search as before:

2. Check the *Search Lemma* option and carry out a new search for *show*. How many total translation units are there?

There are 401 translation units returned for this search because by specifying *lemma*, the search includes all forms of *show*, such as the gerund, the past participle and different verb conjugations. Also noted is the Galnet dictionary entry that appears at the top of the page listing all possible Spanish equivalents according to Galnet. Furthermore, it is the lemma that is highlighted in the annotated text instead of the search word.

For the third exercise, the user is asked to carry out a similar search as the previous one, but now in Spanish:

3. Clear the English search box and type *mostrar* into the Spanish search box (maintaining the *Search Lemma* option). How many total search results are there?

There are 258 total results from this search. The Galnet dictionary entry again appears at the top, this time listing the English equivalents instead of the Spanish equivalents. This exercise is simply to get used to using both languages, as well as to draw the user's attention to the number of results in order to see how the different options affect the search outcome.

Once the user has become familiarized with the basic search options, it is necessary to gain a deeper and more practical understanding of Galnet as this is the underlying foundation for SensoGal:

> 4. How many total results are there when searching for *mostrar* and *show* simultaneously (keeping *Search Lemma* checked)? After filtering out those results without lexical equivalents (via the funnel icon), what is the Galnet concept for the second result (PEA – 715, numbered as 4)?

The search yields 104 results and the lemmas in both languages are highlighted in yellow. This exercise has two objectives: 1) practice bilingual searches in SensoGal, and 2) get the user to explore the basics of Galnet on his or her own. By clicking the ILI code from the second matched result, the user is taken to the Galnet page for the concept *give expression to*. At this point the tutorial lays out the hierarchical structure of Galnet (adopted from WordNet), explaining the concepts of hypernyms and hyponyms. Put simply, in a hierarchical structure such as Galnet and WordNet, hypernyms are general concepts made up of more specific concepts, hyponyms. *Color* is a possible hypernym of *blue*, *red*, *green*, etc, because these are more specific examples of the general idea of "colors". In the case of *show,* the hypernym is *convey*, which corresponds to the concept *make known; pass on, of information*. This makes sense as *showing* is just one way of making something known or passing on information. *Show* can also be a hypernym itself as it has more hyponyms below it in the conceptual hierarchy, such as *express through a scornful smile*. This can also be visualized on the concept's Galnet page, as in Figure 29 below.

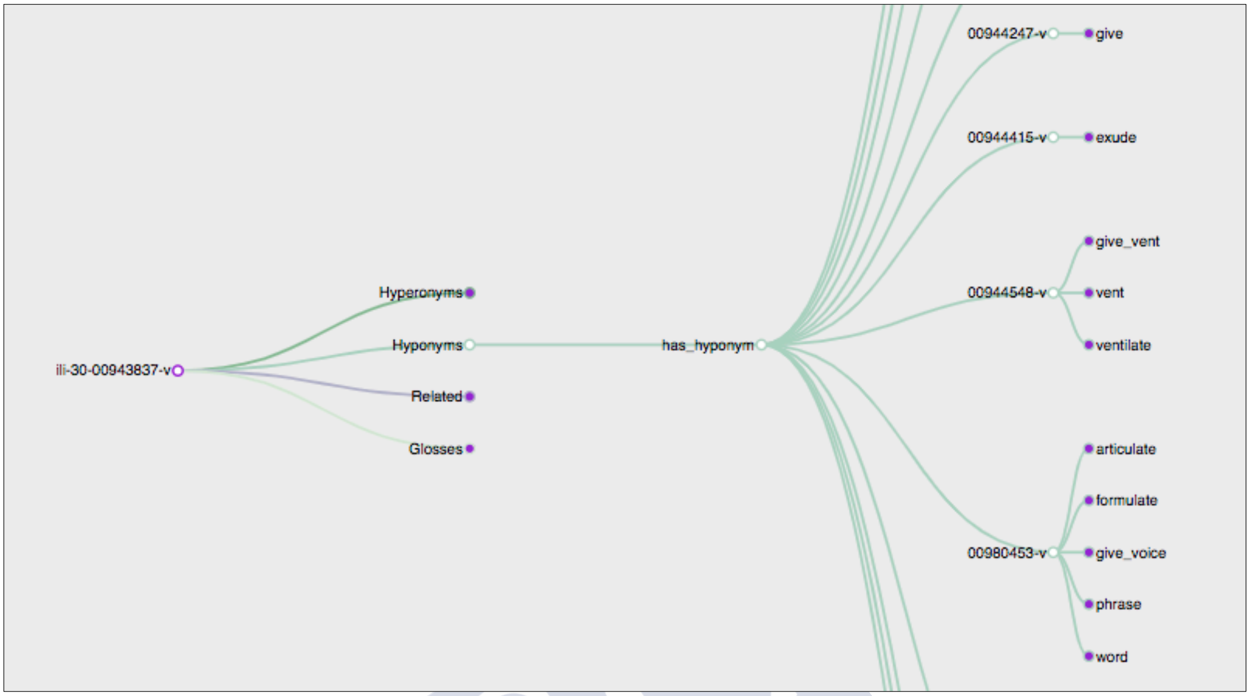


Figure 29. Graphic showing the hyponyms for *give expression to*.

Because Galnet has been adapted from WordNet, a semantic dictionary originally developed in the United States at Princeton University, all sense descriptions are in English.

In order to reinforce the difference between searching by word and searching by lemma, the next exercise asks the user to uncheck the *Search Lemma* option and carry out the same search as in the previous exercise:

5. Repeat the simultaneous search for *mostrar* and *show* but uncheck the *Search Lemma* option. How many total results are there now?

This search will only return 21 results because when searching by word, only the exact words from the search are returned and no other forms; that is, *show* and *mostrar*. It is noted that the words are now highlighted in the text rather than the lemmas beside the ILI code.

Because SensoGal has been automatically tagged, there are inevitably instances where the sense attributed to a word is questionable or clearly an error. That is what the next exercise attempts to make the user aware of:

6. Set the number of results to *All*. Now search for *show* by lemma (leaving the Spanish box empty), but specifying *noun* as the part of speech. What is the ILI code for the only two semantically aligned equivalents?

In order to find the only two semantically aligned equivalents, the user must make use of the funnel option presented earlier on. The only matching semantic equivalent for *show* as a noun is *espectáculo*. Yet, as the video points out, in the original set of results, *espectáculo* appears multiple times in the translation but is not marked as a semantic equivalent because it is not annotated with the same ILI code and therefore corresponds to a different concept. This is meant to remind the user to approach any type of automatic annotation with a critical eye to avoid inaccurate or incomplete results.

Lastly, to demonstrate the Spoken English Variety function the user is told to search for a word that is frequently pronounced differently in American and British varieties of English:

7. Search for *adult* by lemma in English. In which variety of spoken English is the first syllable stressed? In which variety is the second syllable stressed?

The data shows that in British English the first syllable in *adult* is stressed while it is the second syllable that is stressed in American English. It is important to realize that when searching by spoken variety, any possible results from the other variety are not simply hidden, but rather completely omitted from the search results.

Part two concludes with a recap of what was covered in the video tutorial, specifically searching by word, searching by lemma, basic search functions and a cursory glimpse of semantic annotation and Galnet, which will be covered further in part three, Search by Concept.

3.1.2.2 Search by Concept

Now that the user is familiar with the basics of Galnet, part three will expand upon some of the concepts brought up in part two, specifically the idea of lemmatization and semantic annotation, and how the latter corresponds to concepts, or "senses", in Galnet.

In order to search by concept, it is necessary to obtain an ILI code. This can be done by browsing Galnet or by simply locating the ILI code in a particular set of search results. The first exercise begins with the latter:

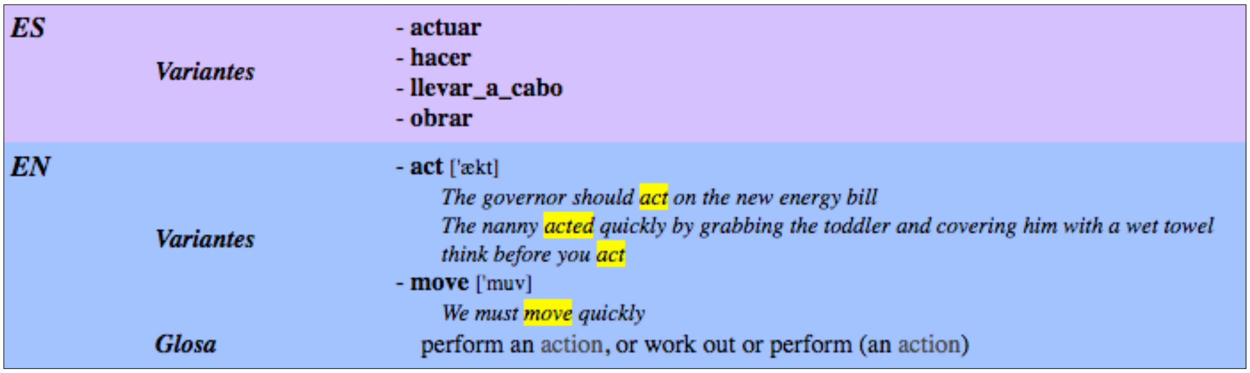
1. Search the word *act* by lemma. What is the ILI code for *act* in the first translation unit?

The ILI code for *act* in the first TU is *ili-30-02367363-v*. It is then explained that each ILI code begins the same way, ili-30-, followed by an identification number, followed by the part of speech (noun – n, verb – v, adjective – a, adverb – r). In this case we can see that the concept corresponds to a verb.

The user is then instructed to click on the ILI code to find additional information in Galnet:

2. Click on the ILI code for *act* in the first translation unit. What is the glossary definition for this concept (in English)? What are the Spanish equivalents for this concept?

The concept is *to perform an action, or work out or perform (an action)*. The Spanish equivalents are *actuar*, *hacer*, *llevar a cabo* and *obrar*.



ES	***Variantes***	- **actuar** - **hacer** - **llevar_a_cabo** - **obrar**
EN	***Variantes***	- **act** ['ækt] *The governor should act on the new energy bill* *The nanny acted quickly by grabbing the toddler and covering him with a wet towel* *think before you act* - **move** ['muv] *We must move quickly*
	Glosa	perform an action, or work out or perform (an action)

Figure 30. English and Spanish Galnet entries for the concept *perform an action, or work out or perform (an action)*.

Next, the user is instructed to carry out a search using an ILI concept:

3. Returning to the search results, copy the ILI code for *act* from the first translation unit and paste it in the English box in the *Search by concept* section of the search menu. How many total results are there?

This search yields 74 results, all returning variations of *act* despite *move* also being listed alongside *act* as an entry for the same concept in Galnet. This is then contrasted with the results when searching for the same concept in Spanish:

4. Now search for the same concept, but only in Spanish. How many total results are there? Why are the results different?

Now there are 208 results. The number of results differs due to the fact that there are more words in Spanish (*actuar*, *hacer*, *llevar a cabo*, *obrar*) that pertain to this concept than words in English (*move* & *act*). This disparity is also due to the fact that the corpus, like all corpora, has its limits and is of a certain size and domain. Therefore, not all instances of all language use can be included in a corpus and such examples of *move* were likely from another corpus where the data found it to correspond to that concept.

The final exercise makes the user aware of the fact that searches for differing concepts may be carried out in both languages simultaneously:

5. The first result from the previous search did not contain semantic equivalents. Search for the concept corresponding to *work* in the first English translation unit (ili-30-02413480-v) while simultaneously searching for the same concept from the previous search (ili-30-02367363-v) in Spanish. How many different books yield results for this search?

The search yields results from three different books: *The Pearl*, *The Call of the Wild* and *1984*. The user can compare the two concepts and their glossary definitions. The concept corresponding to *work* is to *exert oneself by doing mental or physical work for a purpose or out of necessity*. While *work* is the only single entry in English for this concept, there are two words in Spanish, *esforzarse* and *trabajar*. We can deduce from these search results that the two concepts are indeed similar. It is not surprising, then, that two similar concepts may align rather than the same exact concept. This is not a shortcoming of the corpus as much as it is a testament to the richness of language and the complexity of translation, particularly literary translation.

The tutorial concludes with a reminder of caution when working with Natural Language Processing, or NLP, technologies such as automatic semantic tagging as it is still far from an exact science and would require an amount of human resources that are simply not available in order to achieve anything near "perfect".

3.2 SPEECH PRODUCTION AND COMPREHENSION

As stated previously, relatively very little research has been carried out on the study of English phonology through a speech corpus in ESL contexts, which is what has lead to LITTERA's inception. In this section, a series of examples will be provided to show how this can be achieved.

The work below draws upon research in the field of contrastive analysis as it pertains to Spanish and English and will address common phonological aspects of English language curricula. Many of the issues addressed in the following sub-sections are frequent, observable problems in the acquisition of English by Spanish speakers brought on by the phonotactic differences between the two languages.

Two features of spoken language will be regularly brought up in the sub-sections to come: segmentals and suprasegmentals. Segmental features are the individual sounds, i.e. vowels and consonants, while suprasegmental features "involve stress, rhythm, intonation, and coarticulatory phenomena which occur under the influence of stress and intonation, such as elisions, contractions and assimilations" (Chela-Flores, 1997). All these features frequently fall under the umbrella term *prosody*. While it is common in language classrooms to begin with the individual sounds and build upward to lexical chunks and eventually to whole phrases and sentences, many researchers have argued for the opposite approach; that is, to begin with aspects such as rhythm and word stress (Solé Sabater, 1991; Chela-Flores, 1997) as those are the factors that often determine segmental realization, particularly at word boundaries. As Kjellin (1999) puts it, "[a] very important aspect of prosody for the second-language learner is the interconstituent interaction between prosody and the segmentals" and that "segmental mispronunciations will even go completely undetected by the native listener, if the student pronounces the prosody correctly". Kjellin reasons that this is "due to the holistic and anticipatory nature of native perception, i.e. the fact that when the prosody is the expected one, the input signal as a whole will pass through the native listener's phonological filter, as it were".

One cannot talk about segmentals without considering the utterance in which they exist due to the nature of co-articulatory phenomenon in English. English is often referred to as a *stress-timed* language, meaning that English rhythm depends on the number of stressed syllables in an utterance and that there tends to be an even amount of spacing between stressed syllables (Pennington, 1996), regardless of the number of unstressed syllables in between (Whitley, 2002:66). These groups of syllables are referred to as *feet*, and each *foot* contains a stressed syllable along with the other lesser-stressed syllables. An English foot "takes about the same time to pronounce whether it has one or four syllables" (Whitley, ibid; 66). To achieve this even spacing, Whitley explains that syllables are "shortened and squeezed" into the foot to accommodate English rhythm. Compare this to the so-called *syllable-timed* languages, such as Spanish, in which syllables occur fairly evenly, regardless of stress (Carr, 1999).

However, this neat distinction between *stress-timed* and *syllable-timed* languages is not as black and white as it may first appear. Some researchers (Giegerich, 1992; Schlüter, 2009) argue that this categorical distinction is not in line with current research and, in reality, is far more "one-dimensional" and "scalar". Therefore, "a gradient of isochrony[30] is proposed [in the current research], with stress- and syllable-timed languages at either extreme of the continuum" (Schlüter, ibid). On such a continuum, English would gravitate further toward the *stress-timed* end while Spanish would lie closer to the *syllable-timed* end. Regardless of the descriptive details, it is clear throughout the literature that stress and other prosodic features play an important role in English phonology, as they lead to vowel reduction and other co-articulary phenomena that will be explored in the corpus data below.

Even as far back as the 1990s, teachers and researchers were becoming aware of the importance of suprasegmentals in spoken discourse (Jones, 1997). *Unnatural* or *non-native-like* prosody can be one of the main causes for the perception of a foreign accent (Jilka, 2000). Research has also shown that "explicit instruction focusing on suprasegmentals, more so than

[30] *Isochrony* refers to "the regular occurrence of stresses" (Solé Sabater, 1991) that are "roughly equal in time" (Giegerich, 1992).

segmental training, leads to improvements in spontaneous L2 speech" (Trofimovich & Baker, 2007). All this is not to say that segmentals ought to be ignored altogether, but rather that it is necessary to first have an understanding of the various elements of prosody that regularly act upon individual segments in connected speech.

It must be reiterated here that even though intonation may be observed in the corpus, it will not be the focus of any analysis in the present work for two reasons: 1) although attempts have been made to catalogue English intonational patterns (see Stockwell & Bowen, 1965), such tendencies are just that, and also rely heavily on the context in which a given utterance takes place, and 2) the audio in the corpus is not spontaneous speech. Even in professional recordings it is an approximation of how a character from the text *might* say something. It is an interpretation; a kind of guesswork. That's not to say that useful information on intonation cannot be extracted from the corpus data, but rather it is not the focus of the present work.

Much of the features discussed below will be analyzed through the CLUVI corpus, although there are some instances in which the SensoGal version will also be incorporated. Unless otherwise noted, all search queries will be referring to the CLUVI corpus.

Lastly, it should be noted that although the examples below are meant for Spanish speakers, many of the issues pertaining to speech production and comprehension may also be relevant to all learners of English. It is the author's hope that the present work inspires learners of English from all language backgrounds to embrace corpora in the study of English phonology.

3.2.1 Suprasegmental Features of English

The following sub-sections will examine common tendencies of connected English speech found in the corpus data that will be useful for both improved oral comprehension and achieving more native-like speech production, especially since English orthography tells us far less about spoken English than Spanish orthography tells us about spoken Spanish.

In order to describe the co-articulatory phenomenon in the data, it is necessary to introduce some general phonological concepts pertaining to connected speech, specifically *linking*, *assimilation*, *elision*, *palatization*, *weak form* and *tone unit*. *Linking* refers to how words connect in fluid speech, especially in terms of ambisyllabic or resyllabified segments across word boundaries. An example of this is the /k/ in *take out* /teɪ.kaut/. It links the two words with a resyllabified /k/ as it no longer has the sole function of a word-final consonant and also functions as a word-initial consonant. *Assimilation* refers to when one segment takes on features of a neighboring segment. Kreidler (1993) provides an example of assimilation in voicing with the word-final -*s*- in *bets* and *beds*. The /s/ takes on the voicing of the preceding phoneme in *beds* to become /z/. This is an example of *progressive* assimilation. In *regressive* assimilation, the opposite happens—a segment takes on features of the proceeding segment, e.g. /z/ → /s/ in *has to* thanks to the voiceless /t/ in the word-initial position after *has*. *Elision* refers to the omission of a segment, as in the word-final /t/ in *first sight. Palatization* is when /t, d, s, z/ occur in word-final position and /j/ occurs in word-initial position in the following word, causing /t, d, s, z/ to move back in the mouth, often becoming /ʧ, ʤ, ʃ, ʒ/, respectively, as in *won't you* /won.ʧu/ or *did you* /dɪ.ʤu/, for example. *Weak form* refers to when function words (i.e. grammatical words such as articles, pronouns, prepositions, auxiliary verbs, etc.) are pronounced with vowel reduction and/or elision of certain segments. This happens frequently in connected English speech in order to make the content words (i.e. lexical words such as nouns, verbs, adjectives and adverbs) more prominent, avoid complex clusters, and maintain isochrony (the more or less even spacing between feet). An example of this is the reduction of the auxiliary *have*, /hæv/ → /əv/, e.g. in *should* ***have*** *known*. Thanks to vowel reduction, "a syllable that is stressed in a citation form may be unstressed in connected speech" (Giegerich, 1992). Finally, a *tone unit*, sometimes referred to as an *intonation unit*, *intonation phrase, sense group* or *thought group*, refers to "a prosodic unit of speech containing at least one syllable that receives phrasal stress (pitch accent) and ends with a boundary tone" (Kim, 2007), not to be confused with the idea of intonation which refers to the melody and pitch contours of an utterance (Chela-Flores, 2003). Tone units are often described as *chunks* as together they tend to make up a larger utterance. The boundaries are often marked by punctuation in written English, such as commas or periods, although this is not always the case, as tone unit boundaries may also appear as a result of syntactic structures such as nonrestrictive relative clauses, as well as pragmatic and affective factors (Watson & Gibson, 2004; Kim, 2007). Tone units are most frequently set off by a pause or a drop in pitch (Gilbert, 2008), as will be discussed below in section 3.2.1.1.

Despite the fact that LITTERA does not contain spontaneous speech, scripted speech poses certain advantages, especially when studying prosody. As Solé Sabater (1991) points out, "prose read aloud or formal speech is more rhythmical than conversational speech". This allows certain aspects of connected speech to be examined without the "messiness" of unplanned speech from certain factors such as false starts, hesitations or other types of interference. While such interferences are undoubtedly relevant to students from a pragmatic point of view, they are not necessary for an examination of the underlying phonological phenomena that characterize English speech. Their absence, therefore, may provide more accessible data in terms of suprasegmental features of the language.

One last note must be made regarding the data in the corpus and the exercises below. It is necessary to distinguish between categorical rules, i.e. rules that apply to all English speakers, and variable rules, i.e. rules that may vary depending on affective qualities (class, age, sex, dialect, etc.) as well as stylistic qualities, such as how fast one chooses to speak. Therefore, in many cases, it is likely that a certain "rule" will be more of a tendency, or a variable rule, according to the data and not categorical in nature.

3.2.1.1 Locating Tone Unit Boundaries

In order to understand suprasegmental features of English, students must first be aware of tone unit boundaries. Most speakers are aware of them at an unconscious level as they offer cues that help the listener decode the message. Native speakers perceive and understand these cues intuitively. Within each tone unit, certain syllables are stressed (tonic syllables) and thereby certain words—usually, though not necessarily lexical words—are made more prominent. This is essential to effective communication in English:

> "Learners typically do not use or recognize the cues that native listeners count on to help them follow meaning in a conversation. As a result, conversational breakdowns occur. Emphasis that conveys the wrong meaning, or thought groups [i.e. tone units] that either run together or break in inappropriate places, cause extra work for the listener who is trying to follow the speaker's meaning. If the burden becomes too great, the listener simply stops listening. The principle of "helping the listener to follow," therefore, is a vital one. It is so central to communication, in fact, that time spent helping students concentrate on the major rhythmic and melodic signals of English is more important than any other efforts to improve their pronunciation." (Gilbert, 2008)

Learners need not become experts on tone units, but rather made aware of their existence. Gilbert suggests teaching students how to listen for tone units through simple practice exercises with recorded speech. This can be accomplished using the corpus data.

Students can practice locating tone units with any audio file from the corpus. The first thing students can listen for are the prominent words, most frequently lexical words. The prominent words will have lengthened tonic syllables that the student can also listen for. A search for, say, ***talk*** (displaying 1 per text), will allow the students to hear sentences read by a variety of narrators with differing styles. Students can see how speaking style (slow/fast, formal/informal, dramatic/ordinary) has a direct effect on stress, rhythm and the number of tone units. They can compare narrators from the different texts to get a feel for their style and accent. It may be easier to begin with shorter sentences, such as result number 188[31] from *The Cathedral*, *CAT (46), She wanted to talk*, in which there is only one tone unit consisting of the entire sentence. A good example of a short sentence with two tone units is from *Fahrenheit 451*, *FAH (128)*, *But what do you talk about?*. Even though there is no comma, there are two clearly marked tone units, with the boundary after *but*. In this case, there is a significant pause, which makes the separation of tone units rather obvious. In other examples, it might be less obvious where the boundaries lie such as result 81 from *Heart of Darkness*, *DKN (110)*, *I have been in some of them, and…well, we won't talk about that*. In this example, there are three tone units, divided up as follows (with // marking tone unit boundaries):

// I have been in some of them, // and…// well, we won't talk about that. //

If students only go by punctuation, they are likely to exclude *well* from the tone unit, as it is often pronounced separately from the main clause. However, in this example that is not the case.

After locating the boundaries, students can turn their attention to the composition of the tone unit, specifically the stressed words. By drawing attention to prominence, students become aware

[31] When results are referred to by the number found on the TU (not the one in parenthesis), this may or may not correspond to the order in which it appears on the screen due to the fact that when displaying 1, 5, 10 or 15 per text, the other results are only hidden and not technically excluded. For example, the third result visible on the page may be result number 9, as this is the serial number that appears before the dash, e.g. 9 – SDO or 529 – DKN, and not the number in parenthesis, which refers to the TU's index number within that specific text.

of the fact that stress plays an important role in English speech. The stressed words in the previous example are *been*, *some*, *and*, *talk*, and *that*. Perhaps the most surprising stressed word from this example is *and*. However, because it is said with a large pause before and after, it is its own tone unit with a fully realized vowel (see 3.2.1.6 below for more on vowel reduction).

In terms of tempo, students can compare the examples from *The Call of the Wild*, *CAL (1125)*, and *1984*, *NIN (162)*:

> *CAL (1125)*: *// He never forgot a kindly greeting or a cheering word, // and to sit down for a long talk with them // ("gas" // he called it) // was as much his delight as theirs. //*
>
> *NIN (162)*: *// But // at any rate he had the appearance of being a person that you could talk to // if somehow you could cheat the telescreen // and get him alone. //*

In the first example from *The Call of the Wild*, the narrator reads noticeably more slowly, which is why, along with the existence of an extra comma and a parenthetical, the spoken audio lasts almost four seconds longer than the other, despite the fact that they both contain 37 syllables. The more slowly a narrator reads, the more tone units are likely to appear, as is the case with *gas* from *CAL (1125)*. In terms of stress, the contrast between prominent and reduced words is much greater in the second example. Tempo has a significant effect on pronunciation (Kreidler, 1993) and faster speech will lead to greater vowel reduction for unstressed words, as will be discussed in more detail below.

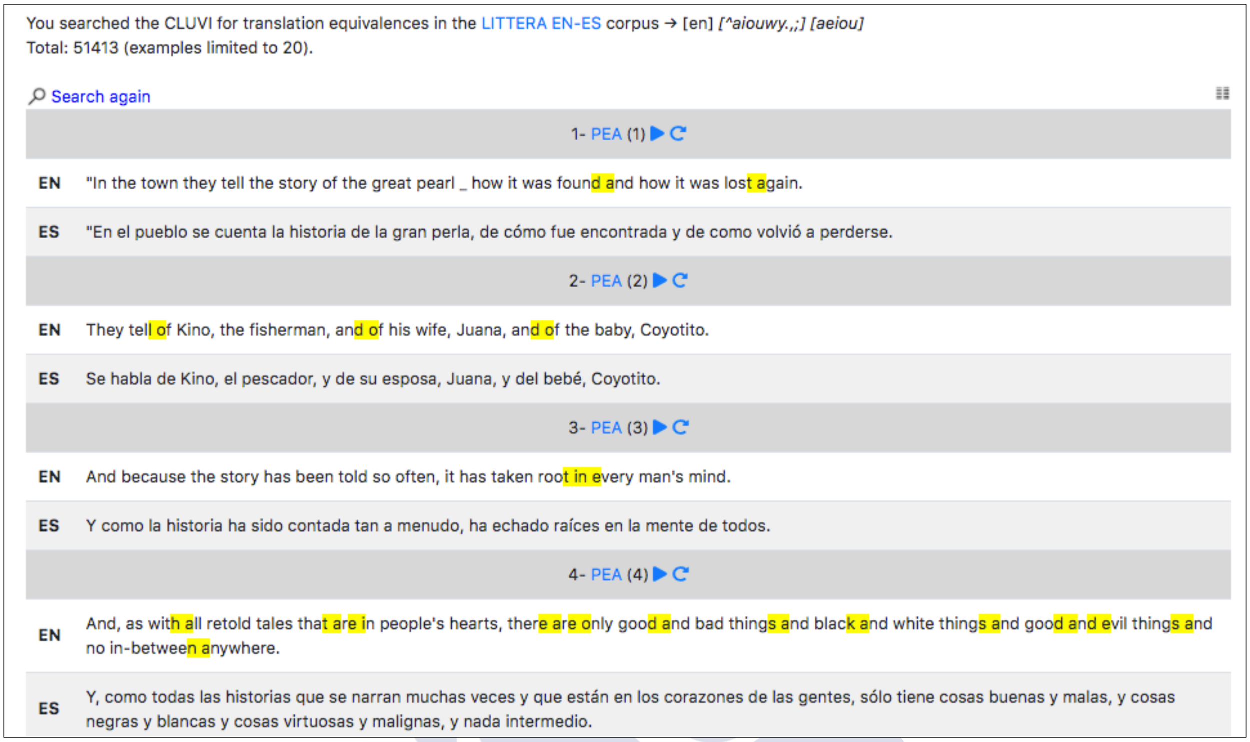
3.2.1.2 Linking across Word Boundaries

As stated above, linking refers to how speakers fluidly move from one word to another in connected speech. Understanding how words are linked together in English is important because the affixes that mark tense or number appear at word boundaries and are affected by the features of prosody. Therefore, learners must be aware of how essential grammatical cues, often spelled with -*s*- and -*d*- (Gilbert, 2008), can be altered by prosody. As Field (2003) points out, English inflectional endings, /t/, /d/ and /s/ "are most subject to assimilation and elision". Linking maintains cohesion within a tone unit and holds it together at the expense of certain segments, often those responsible for marking grammatical cues.

There are four ways words are linked: 1) word-final consonant + word-initial vowel (C+V); 2) word-final consonant + word-initial consonant (C+C); 3) word-final vowel + word-initial consonant (V+C); 4) word-final vowel + word-initial vowel (V+V). The final two ways, V+C and V+V, cause the least amount of change and therefore tend to bring about fewer problems for learners. Therefore, it is the first two, C+V and C+C, that will be addressed in this section. In the case of C+C, only instances in which both consonants are the same phoneme will be considered here to illustrate the phenomenon of blending. The behavior of differing consonants forming clusters across word boundaries will be examined in the following section, 3.2.1.3.

To find instances in which a word-final consonant is followed by a word-initial vowel, C+V, the search query can be formulated as ***[^aiouwy.,;]_[aeiou]*** (recall that the underscore is simply being used here to represent a space in the actual search query). The caret within brackets is exclusionary, and therefore any of those characters will not appear in the results. Also, note that -*e*- was not included due to the fact that it frequently goes unpronounced when in word-final position as it often comes after a consonant in words like *there*, *borne*, *late*, etc. Nevertheless, instances of *the* and *be* followed by a word-initial vowel will still be included in the search query.

Such an incredibly broad search, however, will provide an unmanageable number of results, both in terms of translation units (51,413) and the number of C+V occurrences across word boundaries within the translation units, as can be seen in Figure 31.



You searched the CLUVI for translation equivalences in the LITTERA EN-ES corpus → [en] *[^aiouwy.,;] [aeiou]*
Total: 51413 (examples limited to 20).

1- PEA (1)

EN "In the town they tell the story of the great pearl _ how it was found and how it was lost again.

ES "En el pueblo se cuenta la historia de la gran perla, de cómo fue encontrada y de como volvió a perderse.

2- PEA (2)

EN They tell of Kino, the fisherman, and of his wife, Juana, and of the baby, Coyotito.

ES Se habla de Kino, el pescador, y de su esposa, Juana, y del bebé, Coyotito.

3- PEA (3)

EN And because the story has been told so often, it has taken root in every man's mind.

ES Y como la historia ha sido contada tan a menudo, ha echado raíces en la mente de todos.

4- PEA (4)

EN And, as with all retold tales that are in people's hearts, there are only good and bad things and black and white things and good and evil things and no in-between anywhere.

ES Y, como todas las historias que se narran muchas veces y que están en los corazones de las gentes, sólo tiene cosas buenas y malas, y cosas negras y blancas y cosas virtuosas y malignas, y nada intermedio.

Figure 31. Search results for *[^aiouwy.,;]_[aeiou]***.**

Therefore, a much more practical approach is to search by individual sounds, or groups of similar sounds. To help students perceive word-final consonants, Gilbert (2008) recommends making the distinction between *stops* and *continuants.* Stop sounds are formed by a disruption in the airflow through the mouth and include /b, d, g, p, t, k/. Voiceless stops, /p, k, t/ may produce aspiration, especially when they occur at the beginning of a syllable or before a stressed vowel. This is possible in other positions as well, though normally to a lesser degree (Whitley, 2002). All remaining consonants are continuants or some combination of the two, such as the affricates /ʧ/ or /ʤ/.

Due to the aspiration in syllable-initial positions, word-final voiceless stops are a good starting point to illustrate C+V linking. The aspiration from voiceless stops provides easily identifiable evidence for a phenomenon known as resyllabification. Resyllabification occurs when "a syllable-final consonant attaches itself to the following syllable" (Field, 2003). A simple example of this phenomenon is the syllable-final /n/ in *win* [wɪn] becoming a syllable onset in the word *winner* [wɪ.nɚ]. This follows the Maximal Onset Principle, which states that when a consonant has the potential to function either as a coda for one syllable or the onset of another, the consonant takes the onset position[32] (Carr, 1999). Resyllabification may also occur across word boundaries. Field (ibid) provides the examples of *went in* and *made out*, which, due to resyllabification, may sound like *when tin* and *may doubt*, respectively. These false boundary cues can be problematic for learners. Phrasal verbs exemplify this phenomenon quite well, as the particle in phrasal verbs is stressed due to the fact that it is part of the main verb and therefore operating as part of a lexical item rather than a function word, such as a preposition. As a result, the voiceless stop will link to the stressed particle, which allows the aspiration to be heard more clearly. While it is possible to find evidence for resyllabification in other, more general searches for, say, all three voiceless stops before a vowel, phrasal verbs provide some of the clearest examples. Because the main purpose of this task is to make the learner aware of linking in connected speech, the most sensible approach would be to start with the most discernable instances.

An effective way to search by phrasal verb is to include the full particle in the query. Therefore, a search can be formulated with a voiceless stop, followed by a specific particle. An example of this would be ***ke?_off\b***[33]. This query yields 89 results, with only a very small handful of extraneous results. The examples almost invariably show resyllabification, with cases such as *took off* [tə.ˋkɔf] or *broke off* [brə.ˋkɔf], among others. This resyllabification poses a potential problem for learners as the resyllabified /k/ may cause *off* to be understood as *cough*, especially in the case of *took off*, which could be interpreted as *to cough* despite this being grammatically impossible in the vast majority of cases as it is preceded by a subject pronoun. In result 41, *FAH (482)*, *The others would walk off and leave me talking*, there is greater potential for misinterpretation as it could be perhaps understood as *want cough*, which would be a slightly more grammatically plausible interpretation than the others, however semantically absurd it might be. Understanding these shifting boundaries is essential for learners of English to make sense of connected speech.

Another instance in which a phrasal verb may be misheard with a resyllabified /k/ is when the particle is *up*. The search query ***ke?_up\b*** yields 233 results (displaying 5 per text), the vast majority being phrasal verbs once again. In result number 4, *PEA (857)*, *…a little man with a shy*

[32] According to Carr (1999:75), the "[p]reference for filled, rather than empty, onsets is probably rooted in the nature of our articulatory apparatus and also tied to greater perceptual salience of onset consonants".
[33] Recall that the question mark in a search query means zero or one instances of the previous character, in this case *-e-*.

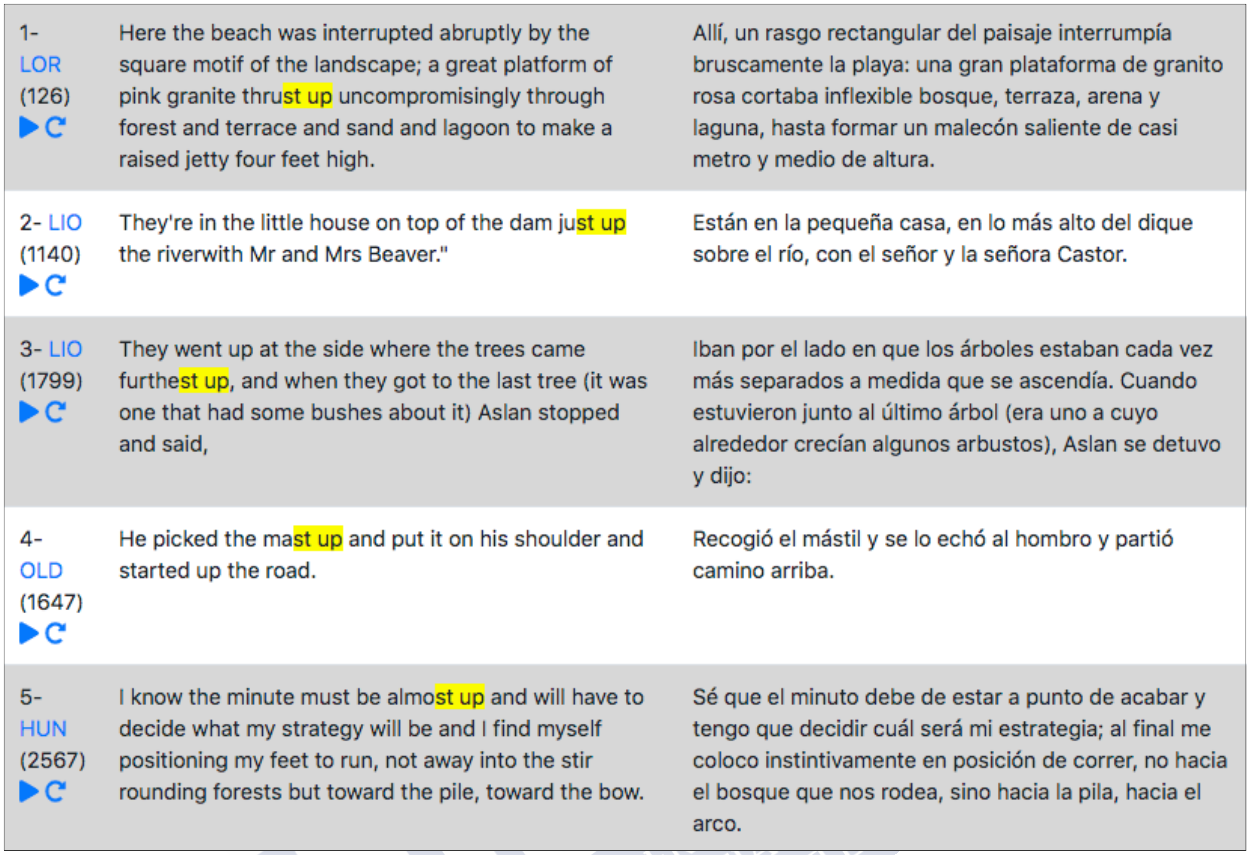
soft voice, took up the pearl…, not only is the /k/ resyllabified, but the lack of any clear grammatical cues could lead to the potential misinterpretation of *to cup*. In result number 9, *SDO (912)*, *…when I make up my accounts in the spring…*, *make up* has the potential to be misunderstood as *may cup*, which would be grammatically acceptable in this context, albeit semantically senseless. The same goes for other examples in the results, such as *back up* (*bad cup**) and *walk up* (*want cup**). In the 18th result, only one of the two results in which there is no phrasal verb, *HOU (69)*, *...and blew little wavering rings of smoke up to the ceiling*, a contrast with the other examples from phrasal verbs can be heard. The resyllabification is less apparent due to the stress pattern of a noun followed by a preposition. Because the two do not form a lexical item, the noun, *smoke*, has a lengthened vowel with a falling intonation, characteristic of a tone unit boundary. Here, word stress provides cues to the listener that *smoke* and *up* are functioning separately.

There are, of course, other word-final sounds beyond voiceless stops that help illustrate resyllabification, such as the voiceless fricative /ʃ/, which is an example of a continuant. The search query ***sh_out\b*** only returns fourteen TU's, ten of which contain phrasal verbs. The resyllabification of /ʃ/ can be misinterpreted as *shout* in examples such as *rush out* (*run shout**) or *gush out* (*gun shout** / *gut shout**). However, in the examples without phrasal verbs, such as *GAT (1040)*, *…all built with a wish out of non-olfactory money*, the stress pattern changes. In this example, much like in the previous example of *smoke up*, the lexical word *wish* is stressed and its vowel lengthened, all with a downward intonation, while *out* is functioning as a preposition and therefore unstressed. Pennington (1996) explains that "[i]n phrases, the unstressed syllables of function words are compressed and blended together in anticipation of the stressed content word, showing clearly that they form a unit consisting of one content word and its pre-modifiers", as evidenced here.

Additional searches can be carried out in a similar fashion in which the word-final consonant is followed by a common phrasal verb particle, such as *on*, *in*, *over*, etc. This type of search also provides students with an efficient method for locating phrasal verbs in the corpus data, especially since particles often add specific semantic nuance regardless of the verb they are paired with, such as *off* indicating separation, division, elimination, etc. Therefore, semantic patterns may emerge when examining phrasal verbs in this way.

It is also possible for entire clusters to become resyllabified. Having two consonants shift instead of one may also help mark this tendency further for students and give them a greater understanding of the flexibility of English syllables in connected speech. Clusters beginning with /s/ are a good way to illustrate this point. A broad search can be formulated as ***s[ptk]_[aeiou]***, although much like in the previous instances, it will be more advantageous to search by phrasal verbs. A search for ***st_out\b*** yields 38 results and allows the user to see a variety of phrasal verbs (*burst out*, *cast out*, *thrust out*) as well as other sequences not pertaining to verb phrases such as *just out* and *beast out*. While the phrasal verbs have the same overwhelming tendency to resyllabify the word-final /st/ cluster as in the previous examples, there is considerable variation throughout the other examples. In result 8, *LOR (2872)*, *…that you might put the* <u>*beast out*</u> *of mind…*, *beast* is stressed and its vowel lengthened, while *out* is unstressed due to the prosody of the sentence. As a result, there is no clear evidence of resyllabification in this instance. However, in result 31, *HUN (2665)*, *…but food will go so* <u>*fast out*</u> *here*, resyllabification is much more apparent. This is the result of word stress within a sentence. In the former, *out* forms part of the tone unit *out of mind*, while in the latter, *out* comes near the end of a tone unit right after the focus word, thus altering its stress and clarity.

A search for ***st_up\b*** will further illustrate the effect of tone unit boundaries on pronunciation and resyllabification. Of the five results, as seen in Figure 32 below, only two are phrasal verbs, *thrust up* and *pick up*, both showing clear resyllabification. In the case of *pick up* from *OLD (1647)*, even though the verb is separated from the particle by the object *the mast*, there is clear resyllabification of /st/ as *up* is still functioning as part of the verb phrase. Of the remaining three, two occur at the end of tone units, *LIO (1799)* and *HUN (2567)*. *LIO (1140)* also shows resyllabification despite not occurring as part of a phrasal verb.



1- LOR (126) ▶ ↻	Here the beach was interrupted abruptly by the square motif of the landscape; a great platform of pink granite thrust up uncompromisingly through forest and terrace and sand and lagoon to make a raised jetty four feet high.	Allí, un rasgo rectangular del paisaje interrumpía bruscamente la playa: una gran plataforma de granito rosa cortaba inflexible bosque, terraza, arena y laguna, hasta formar un malecón saliente de casi metro y medio de altura.
2- LIO (1140) ▶ ↻	They're in the little house on top of the dam just up the riverwith Mr and Mrs Beaver."	Están en la pequeña casa, en lo más alto del dique sobre el río, con el señor y la señora Castor.
3- LIO (1799) ▶ ↻	They went up at the side where the trees came furthest up, and when they got to the last tree (it was one that had some bushes about it) Aslan stopped and said,	Iban por el lado en que los árboles estaban cada vez más separados a medida que se ascendía. Cuando estuvieron junto al último árbol (era uno a cuyo alrededor crecían algunos arbustos), Aslan se detuvo y dijo:
4- OLD (1647) ▶ ↻	He picked the mast up and put it on his shoulder and started up the road.	Recogió el mástil y se lo echó al hombro y partió camino arriba.
5- HUN (2567) ▶ ↻	I know the minute must be almost up and will have to decide what my strategy will be and I find myself positioning my feet to run, not away into the stir rounding forests but toward the pile, toward the bow.	Sé que el minuto debe de estar a punto de acabar y tengo que decidir cuál será mi estrategia; al final me coloco instintivamente en posición de correr, no hacia el bosque que nos rodea, sino hacia la pila, hacia el arco.

Figure 32. Search results for *st_up\b*.

A word-initial vowel following a word-final consonant, C+V, is the simplest way for students to see how words can link together in English. However, another important instance in which linking may not be obvious is when both word-final and word-initial phonemes are the same consonants. Different types of consonants behave in different ways in this situation. For example, "full" continuants (i.e. the sounds that are not stops or affricates) blend the two sounds into one, making the word-final and word-initial sound ambisyllabic. This is normally not problematic for Spanish speakers as the small handful of word-final consonants allowed in Spanish, /n, r, l, ð, s/, are all continuants and display similar blending across word boundaries, e.g. *los santos* [lo.san.tos]. Such blending can easily be searched for in the corpus data by putting the same letter or letters twice and separated by a space, such as ***f_f***, ***sh_sh*** or ***s_s***. For all the continuants, the data shows consistent blending of the word-final and word-initial phonemes. A search for ***s_s*** (displaying 1 per text) is also informative in other ways as instances of assimilation can be observed, usually when the voiced alveolar sibilant /z/ in *was* or *always* takes on the voicing of

/s/, or even /ʃ/, in the proceeding word, becoming voiceless and blended into one sound. An especially illuminating example for students is result number 1801, *HEA (20)*, *…so that I might not disturb the old man's sleep*. Here, the narrator reads very slowly, which one would expect to have the opposite effect. Yet, the blending and assimilation (/z/ → /s/) still take place. This is especially useful information for Spanish speakers as it provides a way to avoid the often problematic *sl-* cluster in word-initial position [34].

The corpus data for affricates, /ʧ/ and /ʤ/ shows just the opposite. There is no blending and the phoneme is repeated. However, the corpus data is limited in this case as a search for ***ch_ch*** yields 15 results and a search for ***dge_j*** yields only one, and the repetition of the phoneme may be the result of scripted speech. Such a relatively low number of results may not reflect the reality of these affricates occurring together in word-final and word-initial position in normal English speech. Fortunately, in terms of comprehension, the corpus data tells us that there is no distortion and, therefore, there should be no impediment to comprehension. The issue, then, would be one of production, as the data provides no way to simplify the juxtaposition of these sounds in connected speech through blending like the continuants above.

Finally, stop consonants tend to neither fully blend like continuants, nor repeat like affricates (except, perhaps, in very slow, deliberate articulation). Instead, it is common for the word-final consonant to be unreleased (marked with a corner diacritic, as in [t̚]), replaced with a glottal stop, [ʔ], or elided altogether. Because the difference between a glottal stop and an unreleased stop may be difficult for the untrained ear to detect, what is most important for students to know in this type of task is that it is unnecessary to pronounce both word-final and word-initial stops as attempting to do so might hinder speech fluidity.

The searches can be carried out in a similar way to the continuants and affricates, although some must be modified to avoid extraneous phonemes from interfering in the results. Therefore, ***g_g*** must be modified as ***[^n]g_g*** to avoid interference from /ŋ/, ***d_d*** modified to ***[^e]d_d*** to avoid interference from the other realizations of the past tense morpheme, ***c_c*** modified to ***[ck]_c[^he]*** to avoid interference from word-initial /ʧ/ and word-initial /s/ as in *chart* or *century* respectively, and ***t_t*** modified to ***t_t[^h]*** to avoid interference from /θ/ and /ð/.

[34] This is due to the tendency of Spanish speakers to put a vowel before s + C clusters as per the phonotactic rules of Spanish (e.g. *esleep** for *sleep*).

In a search for ***[^n]g_g*** (displaying 1 per text), many of the results are simply unreleased stops followed by /g/, however a glottal stop can be heard in the 24th result, *POT (3596)*, *What's that dog guarding?*, realized as [dɒʔ.gɑːdiŋ]. An example in which the word-final stop is elided altogether is the first result from the query ***t_t[^h]*** (displaying 1 per text), *PEA (28)*, *…she went to the…*, resulting in [wɛn.tu]. This result can be compared to another from the same search in which *went to* also appears in the results, as in *HEA (15)*, *I went to work!*. In this instance, the narrator reads very slowly for dramatic effect and fully realizes both the word-final /t/ and the word-initial /t/. However, what is important for students to take away from the data is that there are often easier ways to link words together through blending, resyllabification, assimilation and elision. Therefore, it is not necessary to force the realization of all phonemes in the utterance, as the data shows how infrequent that generally is.

The search tasks in this section have been aimed at making students aware of some of the basic principles of linking in connected speech. Understanding this will equip them with the proper investigative tools to make sense of the phonological data in the corpus. Concepts such as resyllabification and blending explain why certain cues in fluid speech seem to move or disappear altogether. Examples of linking between different consonants acting in clusters can be found in the following section.

3.2.1.3 Complex Consonant Clusters across Word Boundaries

English consonant clusters pose many difficulties for Spanish speakers due to the phonotactic restraints of Spanish. In the previous section, the word-final cluster /st/ before a word-initial vowel was used to exemplify the phenomenon of resyllabification, e.g. *burst out*. This section aims to introduce students to the idea of cluster reduction through elision across word boundaries. As in the previous section, the idea is to provide students with a general awareness of this phenomenon through specific examples that they can then apply to other data they come across in the corpus.

Because Spanish does not allow for word-final consonant clusters, except for a handful of loan words, this feature of English often presents certain challenges for Spanish speakers. As a result, even VCC word-final clusters have the potential to cause difficulty. Complicating this issue further is how these clusters behave when taking into account the surrounding segments.

Because written English contains many silent or double letters, it can be difficult to locate consonant clusters by using an orthography-based search query. An example of this is the word *light*, which has three word-final consonants orthographically, but only one word-final consonant phoneme. This is undoubtedly a clear limitation of the corpus, especially one that aims to examine phonetic data. Therefore, general search queries, like those at the beginning of the last section for any word-final consonants before word-initial vowels, would not be practical for English consonant clusters.

Furthermore, not all clusters are problematic. Much depends on the phonotactic restraints between the two languages, along with the specific learning context in which the corpus is employed. As with most DDL activities, it is necessary for teachers to guide students toward the right data. In the case of consonant clusters across word boundaries, it would make sense to focus on those that *are* problematic and find out whether there are ways to simplify them, which is what the following aims to do.

Given that there a numerous possible combinations of clusters across word boundaries, those analyzed in this section will all be *st* + *C* clusters for the sake of practicality. The word-initial consonants will be broken into the same three groups as in the previous section; that is, stops, continuants and affricates. While there are certainly a fair number of other complex clusters not involving /st/ that may occur across word boundaries, such an exhaustive investigation would go beyond the scope of the present work. The idea here is to show how students can be made aware of when it is possible to reduce clusters according to the corpus data. This will be useful not only for their own comprehension, but also for their own speech production. The following method for searching and analyzing clusters across word boundaries can and ought to be applied to other clusters as well.

All *st* + *C* clusters can be approached the same way, by asking students if it is possible to simplify them in connected speech according to the corpus data. Although batch searches are possible here, individual searches for each specific cluster will provide a much clearer picture for students by allowing them to focus on one specific case at a time. All individual searches can be carried out the same basic way, ***st_C*** (e.g. ***st_b*** or ***st_m***), although in certain cases it is necessary to use regular expressions to avoid unwanted sounds. For example, to obtain results for /k/ in word-initial position, the query could be formed as ***st_[ckq][^h]***, with -*h*- excluded (by using the caret) to prevent the voiceless affricate /ʧ/ from appearing in the results. However, this does not prevent words like *certainly* or *city*, which have a word-initial /s/, from appearing in the results, or words like *know* with a word-initial /n/. To make sure all words beginning with *kn-, ce-* and *ci-* are excluded, we must reformulate the query to ***st_[kg][^n]|st_c[^hei]***. The challenge of formulating queries to increase the accuracy of the results, while likely frustrating to less experienced corpus users, provides a way for students to gain insight into the relationship between English orthography and English phonology.

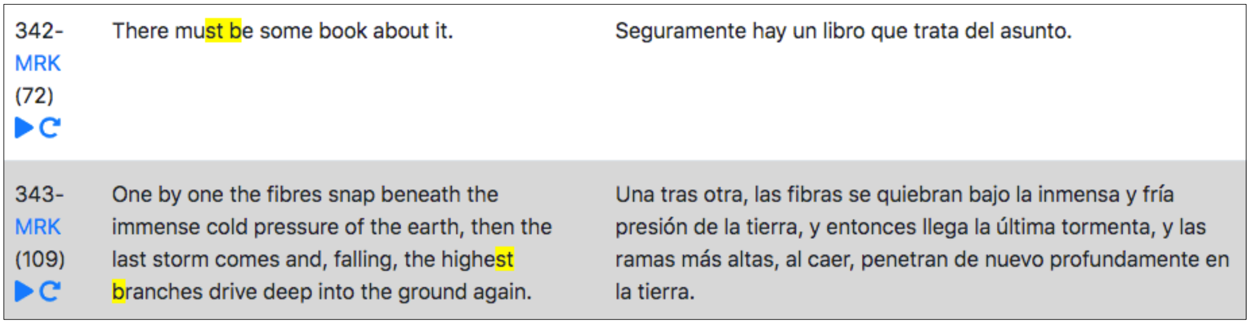
If the initial question is formulated as *Is it possible to reduce 'st + C' clusters across word boundaries?*, then the data is quite clear. All stops and all continuants show the possibility to reduce the cluster, regardless of voicing. The next question could then be *If so, how is this achieved?*. Reduction is achieved through the elision of the word-final /t/ that functions as the middle consonant in the cluster. The results for the query ***st_s*** (limiting the results to 1 per text), for example, show very clear instances of this, although it is not uniform. In results numbers 125, *HOU (6)*, *It was just such a stick…* and 151, *CAL (178)*, *At the first step upon…*, there is obvious blending of /s/ due to the absence of the middle /t/, leading to [dʒə.sʌtʃ] and [fɚ.stɛp], respectively, and occurring in both varieties of English. The same can be observed in a search for *st + d* with a similar query, ***st_d*** (limiting the results to 1 per text). While no pattern between American and British varieties emerges, what is interesting is the comparison between those narrators that did not elide the /t/ in the previous search for *st + s*, but did so in this one. Such is the case with the British narrator for *Sense & Sensibility*. In the previous search, *must say* shows no elision while *most dear* does. Expanding these searches to 5 per text, the narrator seems to maintain this pattern in the other four results, eliding the middle consonant in the *st + d* cluster, but fully realizing the *st + t* cluster. What a student can take away from this data is that the possibility to elide the middle *-t-* exists, but not all speakers do it and this decision may be one of regional accent and/or the result of personal preference.

The data for affricates also show a tendency toward cluster reduction, especially for the voiceless /ʧ/. In fact, in only one of the 13 displayed results (displaying 1 per text) for ***st_ch[^r]*** can the full cluster be heard: *FAH (202)*, *…his chest chopped down…*. Similarly, in the case of ***st_j*** (displaying 1 per text), the full cluster can only be heard in three of the ten displayed results.

Once this tendency is established, the next question should concern the factors leading to reduction, e.g. *What factors may affect the reduction of the cluster?* or *What leads a speaker to reduce a cluster and what leads a speaker to maintain it?*. Students can try to locate examples in which they can point to specific reasons for reduction or lack thereof, such as the presence of a tone unit boundary, which is a common factor in non-reduction. This is not surprising given that word-final segments are often fully realized at a tone unit boundary, as mentioned above. In a search for ***st_b***, a tone unit boundary is the most probable cause for the full realization in *CAL (144)*, *Of this last Buck was…*. The introductory phrase *Of this last*, which when the context is widened is referencing the last item in a list, is a tone unit in itself. Furthermore, the noun and only lexical word, *last*, is the most important element in the phrase and therefore receives the main stress, leading to greater prominence and a lengthened vowel. Another example of a tone unit boundary leading to the full realization of the cluster is in *DUB (95)*, *…the houses that looked to the west reflected…*. As is sometimes the case and evidenced here, a tone boundary may appear at the end of a relative clause. Therefore, the presence or absence of a tone unit boundary must be taken into account as a factor affecting cluster reduction.

Another factor affecting cluster reduction is speech tempo, which is related to tone units, as slower speech often produces more tone units. Because the corpus data consists of scripted speech, it is likely this has some effect on how carefully the narrators speak. The narrator for *The Cask of Amontillado*, for example, shows extremely "clean" or "polished" articulation, which runs the risk of confirming those student biases toward excessively "correct" pronunciation. The narrator consistently maintains the middle /t/ in the cluster throughout all the different searches.

An effective way to examine how tempo affects cluster reduction is to extend the results to 5 per text and look for instances in which the same narrator's speech is either slower or faster. Examples from *The Mark on the Wall* demonstrate this phenomenon. In a search for ***st_b*** (displaying 5 per text), the only two results from this text are *MRK (72)* and *MRK (109)* as seen in Figure 33:



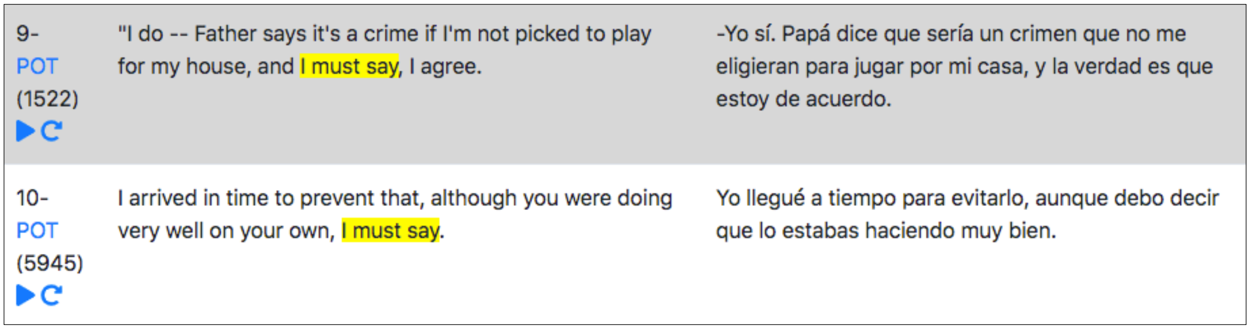
342- MRK (72) ▶ ↻	There must be some book about it.	Seguramente hay un libro que trata del asunto.
343- MRK (109) ▶ ↻	One by one the fibres snap beneath the immense cold pressure of the earth, then the last storm comes and, falling, the highest branches drive deep into the ground again.	Una tras otra, las fibras se quiebran bajo la inmensa y fría presión de la tierra, y entonces llega la última tormenta, y las ramas más altas, al caer, penetran de nuevo profundamente en la tierra.

Figure 33. The only two search results for *The Mark on the Wall* in the query *st_b*.

The first example is read at a quick pace and the cluster is reduced, while in the second example the text is read slowly, most likely for dramatic effect. Not only is the cluster between *highest branches* fully pronounced, so is the cluster in *last storm* from the same TU. A slow tempo provides speakers with time to fully articulate certain phonemes that would otherwise be reduced, assimilated or elided in order to fit them into the rhythm of quick, connected speech.

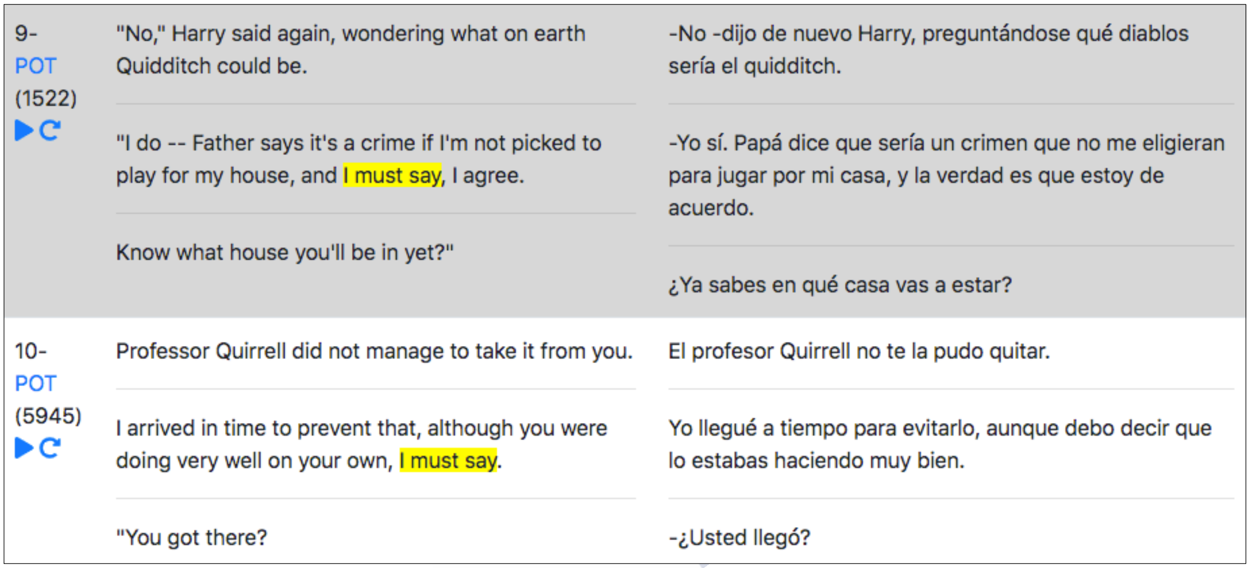
A third factor influencing cluster reduction across word boundaries is the presence of high-frequency phrases or expressions. Because the word-final letters are set to -*st*, it makes sense that high-frequency words like *just*, *almost*, *first*, *last*, *most* and *must* consistently appear in the different searches for *st* + *C* clusters. Although the word endings show varying degrees of realization throughout the corpus, what is particularly relevant to students is the occurrence of formulaic expressions and regular collocations involving some of these high-frequency items.

In the case of the query ***st_s***, the expression *I must say* appears twice in the results (set at 1 per text). In both cases, the cluster is fully realized. Students can carry out a separate search on this expression to confirm this tendency. The 11 results from a separate search show that the cluster reduction via elision of the middle /t/ is, in fact, possible. In nearly all cases in which there are multiple instances from the same narrator, there is consistency in terms of how the cluster is handled, either reduced or fully realized. However, in results 9 and 10, *POT (1522)* and *POT (5945)*, the same narrator gives two different pronunciations. In the first, the cluster is reduced and the middle /t/ elided, while in the second there is no reduction. Initially, it is difficult to say exactly why this happened, but by extending the context it may be possible to get an idea.



9- POT (1522) ▶ ↻	"I do -- Father says it's a crime if I'm not picked to play for my house, and I must say, I agree.	-Yo sí. Papá dice que sería un crimen que no me eligieran para jugar por mi casa, y la verdad es que estoy de acuerdo.
10- POT (5945) ▶ ↻	I arrived in time to prevent that, although you were doing very well on your own, I must say.	Yo llegué a tiempo para evitarlo, aunque debo decir que lo estabas haciendo muy bien.

Figure 34. Results *POT (1522)* and *POT (5945)* without any additional context.

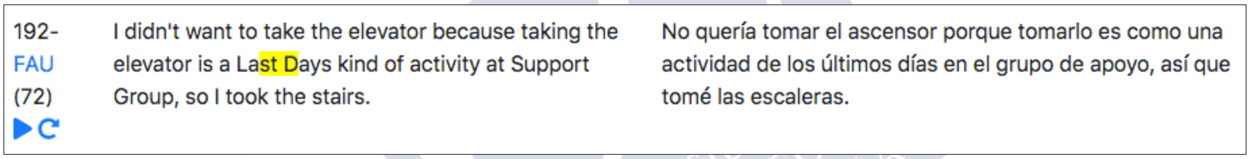


9- POT (1522) ▶ ↻	"No," Harry said again, wondering what on earth Quidditch could be. "I do -- Father says it's a crime if I'm not picked to play for my house, and I must say, I agree. Know what house you'll be in yet?"	-No -dijo de nuevo Harry, preguntándose qué diablos sería el quidditch. -Yo sí. Papá dice que sería un crimen que no me eligieran para jugar por mi casa, y la verdad es que estoy de acuerdo. ¿Ya sabes en qué casa vas a estar?
10- POT (5945) ▶ ↻	Professor Quirrell did not manage to take it from you. I arrived in time to prevent that, although you were doing very well on your own, I must say. "You got there?	El profesor Quirrell no te la pudo quitar. Yo llegué a tiempo para evitarlo, aunque debo decir que lo estabas haciendo muy bien. -¿Usted llegó?

Figure 35. Results *POT (1522)* and *POT (5945)* displaying the wider context.

In the first example, it is clear even without widening the context that it is most likely a child speaking considering the use of the word *Father* as well as *Papá* in the translation, coupled with the fact that the text is *Harry Potter and the Philosopher's Stone* (henceforth *Harry Potter*), a story about a boy wizard. The extended context seems to confirm this supposition, as the speaker is likely a classmate of Harry's. In the second example, *…you were doing very well on your own, I must say* sounds as though a teacher or coach speaking to a student or someone in a submissive or passive role. Furthermore, the Spanish translation also offers insight into this possibility, as the person speaking in the original text uses the familiar *tú* form, and the response, likely from a student, or even Harry himself, uses the formal *usted* form in the response. Therefore, the narrator may be using more careful or polished speech for those in positions of authority and more relaxed or colloquial speech for the younger characters in the text. Admittedly, without a more thorough analysis of the actual narration of that specific text, this can only be considered speculation. This could, however, prompt such questions that may then function as a springboard for more exploratory corpus work.

Other high-frequency items can be found in other searches for the *st* + C cluster. A search for ***st_d*** (displaying 1 per text) yields instances of *day* following *first* or *last*, and even more so if the search is expanded to 5 per text. These collocations can be searched for on their own with the query ***\bfirst_day|\blast_day***. Of the 38 total results, all but one show elision of the middle /t/. The only instance in which the cluster is fully realized is in result number 20, *FAU (72)*, seen in Figure 36 below.



192- FAU (72) ▶ C	I didn't want to take the elevator because taking the elevator is a Last Days kind of activity at Support Group, so I took the stairs.	No quería tomar el ascensor porque tomarlo es como una actividad de los últimos días en el grupo de apoyo, así que tomé las escaleras.

Figure 36. Only result from the query *st_d* to not show cluster reduction through elision of /t/.

The lack of an elided /t/ is not surprising here because *Last Days* is treated as a proper noun and given extra emphasis within sentence, which is not how the sequence *last days* is treated in rest of the data. This falls in line with observations made by researchers that high-frequency phrases, such as *I don't know*, tend to undergo far greater reduction overall than their low-frequency counterparts (Field 2003; Bybee 2006). Some, such as Field, consider it likely that this is due to native speakers storing such phrases as "a single semantic and phonetic entity".

The phrase *in the first place* appears in two separate texts in the search results for ***st_p***. When searched for on its own, seven of the eight narrators reduced the cluster by eliding /t/. The four results from the one narrator (from *The Fault in Our Stars*) who did not were all consistent in their full realization of the cluster. There is no clear evidence to indicate the reason for this other than the preference of the narrator, and a search for this same cluster within the text confirms the tendency of the narrator to fully realize this cluster in nearly all cases.

Finally, the results for ***st_n*** show that the phrase *last night* is repeated four times. The cluster is reduced in all cases, even those read slowly, as in result number 97, *FAH (315)*, *"Last night," he began*. The results from a separate search for *last night* (displaying 1 per text) yield consistent reduction by all fourteen narrators. Besides being a high-frequency lexical bundle in English, it also makes sense that this cluster is normally reduced since the movement in the mouth would complicate the realization of /t/ in preparation for /n/.

Such predictive factors as phonemic class (i.e. stops, affricates, continuants) and frequency can be used to prepare classroom activities. For example, students can be shown a line from the corpus and then told to predict what they think the narrator might do (of course, if students are familiar with the corpus already, then the narrator should be kept anonymous until the recording is actually played). Another possible activity is to have students find examples of a certain phenomenon in the corpus, e.g. *Find five instances in which the consonant cluster 'st + C' is reduced (do not repeat consonants for C)*. These types of 'linguistic hunts' force the listening repetition of the segments in question, which could carry over into speech production with the right guidance on the part of the teacher. Students can also do "reenactment" activities in which they must imitate the speaker (in tempo, rhythm, pronunciation, co-articulation, etc.) to the best of their ability. Lastly, like so many of the other activities here, the corpus can be used for dictation to test learners' ears when it comes to the clusters studied in class.

The initial question proposed at the beginning of this section was one that the corpus could answer fairly easily; that *yes*, it is possible to reduce or "simplify" *st* + *C* clusters across word boundaries. While undoubtedly useful for learners, the question quickly became, *why only sometimes?* Because the corpus data was not uniform, further analysis was done as to what may lead one to reduce a cluster and what may lead one to fully realize it. The possible factors included tone unit boundaries, the nature of the cluster (i.e. which phonemes are involved), speech tempo, a narrator's interpretation of the text and high-frequency items.

3.2.1.4 Assimilation across Word Boundaries

As stated above, assimilation is when one phoneme adopts certain characteristics of a neighboring phoneme, usually in terms of voicing or place of articulation. Recall the voicing assimilation of /z/ to /s/ in the results for the query ***s_s*** in section 3.2.1.2 above. *Man's sleep* was realized as [mæns.slip] instead of [mænz.slip], and the word-final /z/ in words like *was* and *always* became voiceless when preceding a word-initial /s/. In this section we will explore this phenomenon in depth by examining word-final /z/ before /ʃ/, the devoicing of *have* and *has* in the modal verb *have/has to*, the phenomenon of palatization before /j/ (i.e. [wʊd.ju] → [wʊ.dʒu] in *would you*), and the assimilation of *the* to the proceeding phoneme.

A great way to illustrate the idea of assimilation to students is with the high-frequency verbs *is*, *was* and *does* before /ʃ/, even if they are only functioning as auxiliaries. A very common example of this is before the personal pronoun *she*. When assimilated, /z/ becomes /ʃ/ and /ʃ/ becomes ambisyllabic through blending. This activity can be prompted with two questions, such as *How do the speakers in the corpus realize 'is she', 'was she' and 'does she'? Are all the expected sounds fully present?* This prompt can of course be altered to fit the other examples analyzed below.

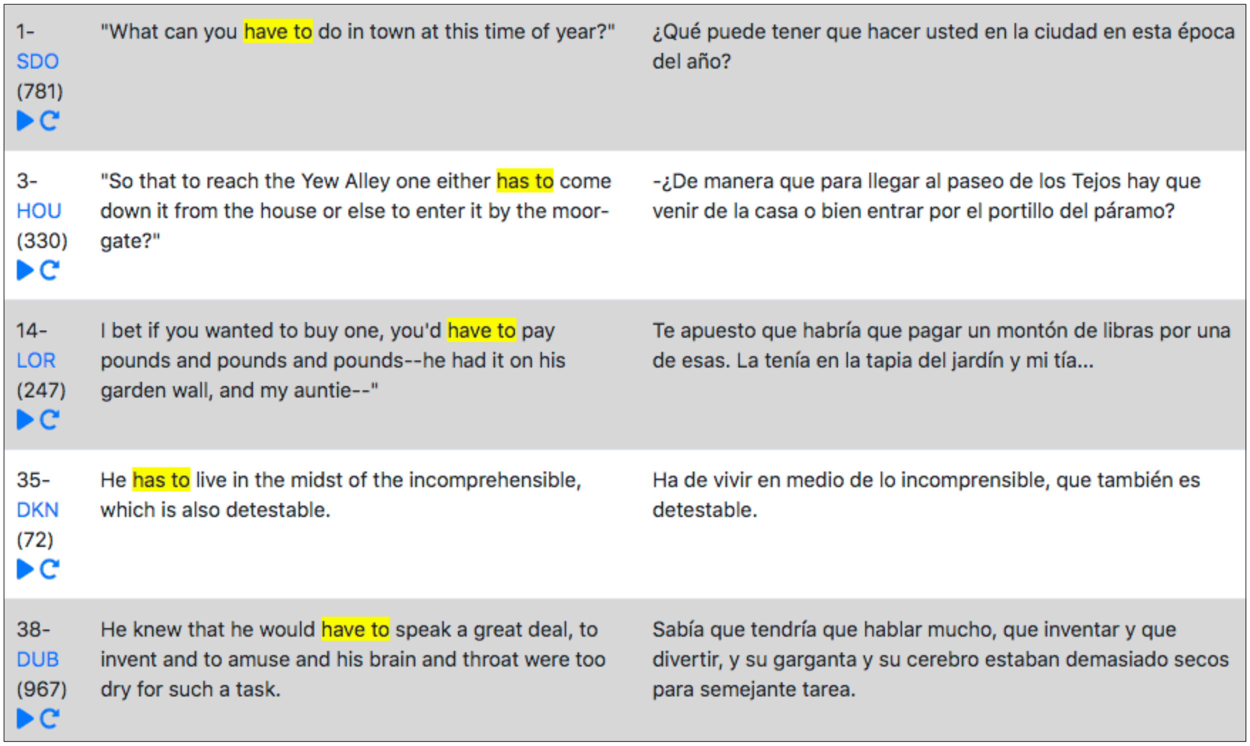
The query for *is she* can be formulated as ***\bis_she\b*** (displaying all results). All 24 results from this search show regressive assimilation, [ˋɪ.ʃi], as it is the preceding phoneme that is altered. Students may note the brevity of many of the sentences in the results. This is because *is she* frequently appears within dialogue, usually in brief questions, e.g. *Who is she?* or *Is she gone?*. The final result, *DAL (647)*, *"And who is she," she asked?*, shows assimilation, even though it is spoken slowly.

A search can then be carried out for ***\bis_he\b*** (displaying 1 per text) to highlight the contrast between the two, mainly those instances in which the initial /h/ is elided, becoming [ˋɪ.zi]. This can be heard in result 20, *DUB (24)*, *Is he dead?*. While some narrators pronounce the word-initial /h/, the examples in which it is elided show how the pronunciation of /z/ at the end of *is* sounds compared to /ʃ/. In fact, these can be searched simultaneously with the query ***\bis_s?he\b*** (displaying 1 per text) and the results provide a fairly even picture, with *is she* appearing five times and *is he* appearing seven.

An examination of *was she* and *does she* can be carried out the same way. Both searches show uniform assimilation of /z/ to /ʃ/, which can then be contrasted with *was he* and *does he* to the same effect as *is he*. The information from these searches is vital for students' comprehension as the cues they normally look for may disappear due to the influence of assimilation.

Furthermore, this is an easy adjustment for students to make in their own speech, in so much as they are able to pronounce /ʃ/.

Another example of regressive assimilation is the case of *have to* and *has to*. *Have to* (in all its forms) functions as a modal verb and is highly frequent in English. In its phonetic realization, the word-final /v/ and /z/ assimilate in voicing to the voiceless word-initial /t/, becoming [hæf.tu] or [hæf.tə] and [hæs.tu] or [hæs.tə]. Much like the devoicing of /z/ in *is/was/does she*, this devoicing is categorical and is evidenced just as overwhelmingly in the corpus data. The search query can be formulated as ***\bhave_to\b|\bhas_to\b*** or ***\bhas?v?e?_to\b***[35], or the two forms can be searched independently.



1- SDO (781) ▶ ↻	"What can you have to do in town at this time of year?"	¿Qué puede tener que hacer usted en la ciudad en esta época del año?
3- HOU (330) ▶ ↻	"So that to reach the Yew Alley one either has to come down it from the house or else to enter it by the moor-gate?"	-¿De manera que para llegar al paseo de los Tejos hay que venir de la casa o bien entrar por el portillo del páramo?
14- LOR (247) ▶ ↻	I bet if you wanted to buy one, you'd have to pay pounds and pounds and pounds--he had it on his garden wall, and my auntie--"	Te apuesto que habría que pagar un montón de libras por una de esas. La tenía en la tapia del jardín y mi tía...
35- DKN (72) ▶ ↻	He has to live in the midst of the incomprehensible, which is also detestable.	Ha de vivir en medio de lo incomprensible, que también es detestable.
38- DUB (967) ▶ ↻	He knew that he would have to speak a great deal, to invent and to amuse and his brain and throat were too dry for such a task.	Sabía que tendría que hablar mucho, que inventar y que divertir, y su garganta y su cerebro estaban demasiado secos para semejante tarea.

Figure 37. Search results for *\bhave_to\b|\bhas_to\b* or *\bhas?v?e?_to\b*.

The searches can also be contrasted with another search in which *have* and *has* are followed by a vowel, such as the indefinite article *a*. This search query can be formulated with the same regular expressions as the previous example, either by using only the vertical pipe (***\bhave_to\b|\bhave a\b|\bhas_to\b|\bhas_a\b***) or the vertical pipe combined with question marks

[35] The two different queries will return the exact same results. These are two different paths to the same end.

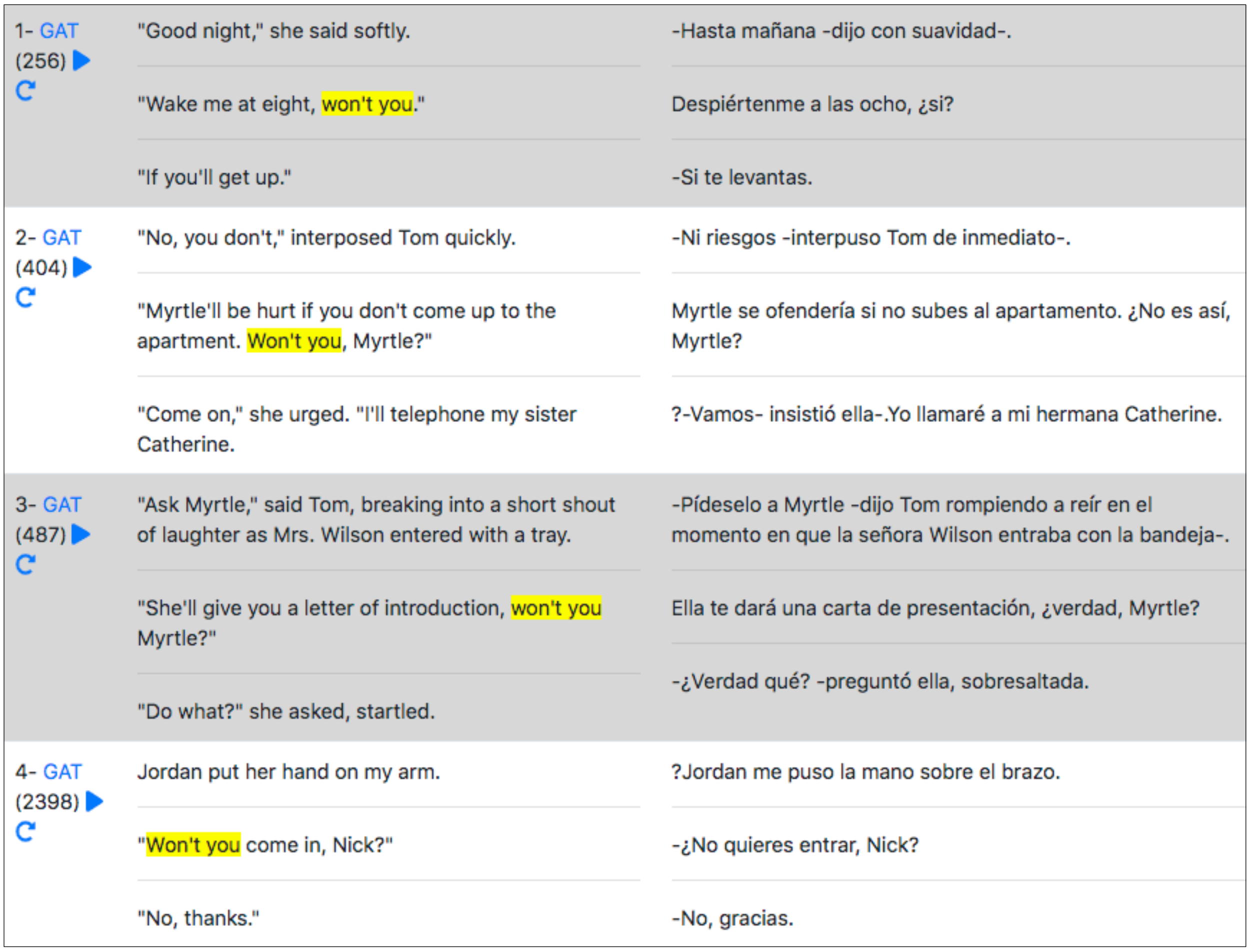
(***\bhas?v?e?_to\b|\bhas?v?e?_a\b***). However, combining these searches and viewing the results as 1 or even 5 per text paints a very skewed picture in terms of frequency, as only 6 out of 84 results when set to 5 per text are a form of *have to*. Therefore, it may be more advantageous for students to search these separately.

Another approach is to search for the forms of *have* and *has* at the end of a tone unit, particularly at the end of a sentence, in order to hear the full pronunciation of the word. This can be carried out the same way, only substituting *a* with the desired punctuation marks. The query for all instances at the end of a sentence can be formulated as ***\bhas?v?e?_to\b|\bhas?v?e?\.\b*** (the slash making the period literal as it is otherwise a regular expression). However, due to the relative infrequency of *have to* or *has to* at the end of a sentence, students may be once again better off searching them separately.

A third example of assimilation that students should be aware of is the palatization before /j/ which occurs when word-final /t, d, s, z/ appear before a word-initial /j/ and is an example of mutual assimilation, meaning that both phonemes are affected, not just one. In English, palatization from /j/ changes the place and manner of articulation, causing /t, d, s, z/ to become [tʃ, dʒ, ʃ, ʒ] respectively. The tongue-tip consonants move backward toward the palate and /j/ moves forward, meeting each other somewhere in between, hence the term *mutual* assimilation. This process is optional across word boundaries and most regularly occurs in high-frequency word combinations (Kreidler, 1993).

Examples of palatization can be searched for by placing *t*, *d*, *s*, or *z* before *y*. In a search for ***t_y*** (displaying 5 per text), clear instances of palatization, /t/ + /j/ to /tʃ/, can be heard by almost all the narrators, namely in *pas**t y**ears*, *bu**t y**ou*, *tha**t y**ou*, *aren'**t y**ou* and *wha**t y**ou*, among others. However, there is no way to fully predict which examples will show palatization. Because some word combinations show a greater likelihood than others or are simply more frequent, such as *but you*, *that you*, *what you*, and contractions ending in *'t* before *you*, students should be encouraged to explore those individually. For example, a search for ***'t_you*** (displaying 1 per text) yields results from 14 different narrators, only four of which do not show assimilation, one of those (*DKN, 696*) being the result of misreading *don't you see* as *you don't see*. These results can be expanded to five per text to find out if those narrators show consistency in their lack of assimilation. An example of this is the narrator from *The Great Gatsby*. Expanding the results for

't_you to five per text, the narrator only assimilates one instance, *GAT (256)*, *Wake me up at eight, won't you* as [won.tʃu]. To see if there is any type of reason for this, students can search within the audiobook itself. The same search for ***'t_you*** in *The Great Gatsby* yields 22 results with an overwhelming tendency towards no mutual assimilation. However, *won't you* appears four times in the results and undergoes mutual assimilation twice. A specific search for *won't you* within the same text and with the context widened reveals that, interestingly, the two instances that show mutual assimilation (*256 & 2398*) are when female characters are speaking, while the other two instances (*404 & 487*) are when a male character is speaking. Furthermore, in the two instances in which the male character is speaking, *you* is reduced to the colloquial form of *ya* [jə], while the other two examples show no vowel reduction for *you* [won.tʃu].



1- GAT (256) ▶ ↻	"Good night," she said softly.	-Hasta mañana -dijo con suavidad-.
	"Wake me at eight, won't you."	Despiértenme a las ocho, ¿si?
	"If you'll get up."	-Si te levantas.
2- GAT (404) ▶ ↻	"No, you don't," interposed Tom quickly.	-Ni riesgos -interpuso Tom de inmediato-.
	"Myrtle'll be hurt if you don't come up to the apartment. Won't you, Myrtle?"	Myrtle se ofendería si no subes al apartamento. ¿No es así, Myrtle?
	"Come on," she urged. "I'll telephone my sister Catherine.	?-Vamos- insistió ella-.Yo llamaré a mi hermana Catherine.
3- GAT (487) ▶ ↻	"Ask Myrtle," said Tom, breaking into a short shout of laughter as Mrs. Wilson entered with a tray.	-Pídeselo a Myrtle -dijo Tom rompiendo a reír en el momento en que la señora Wilson entraba con la bandeja-.
	"She'll give you a letter of introduction, won't you Myrtle?"	Ella te dará una carta de presentación, ¿verdad, Myrtle?
	"Do what?" she asked, startled.	-¿Verdad qué? -preguntó ella, sobresaltada.
4- GAT (2398) ▶ ↻	Jordan put her hand on my arm.	?Jordan me puso la mano sobre el brazo.
	"Won't you come in, Nick?"	-¿No quieres entrar, Nick?
	"No, thanks."	-No, gracias.

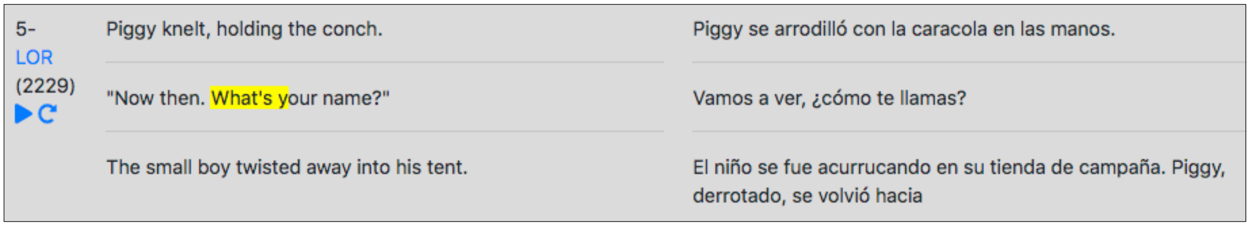
Figure 38. Search results for *'t_you* in *The Great Gatsby*.

This likely deliberate reduction may have also affected the possibility of palatization since the narrator does not avoid it in the other instances. It is possible, then, that this is the result of the narrator's interpretation of speech, specifically what he believes male and female speech is like, or how certain characters might speak. Thus, it would appear that, not only is this a conscious decision, but that it may also have a kind of pragmatic function based on the context of the utterance. *Why might the narrator do this?* is a thought-provoking question for a classroom and is raised as an example of how corpus-driven activities, which from the outset seem simple and banal (e.g. locating instances of palatization), can turn into an examination of a narrator's specific and likely deliberate use of palatization in a text.

Next, a search (displaying 1 per text) for ***de?_y*** returns few results that provide examples of palatization, although one very illustrative example can be heard in *HEA (43)*, *…told you that…*. In this case, the narrator reads slowly, allowing the palatization to be heard clearly: [tol.dʒu]. Because a broad search does not seem to find clear examples, it may be better to have students consider more high-frequency combinations directly, such as *did you*, *would you* or *told you*. In a search (displaying 1 per text) for *did you*, only six of the 17 narrators show no palatization. In the case of *would you*, only two of the 15 narrators show no palatization, and six out of 15 for *told you*.

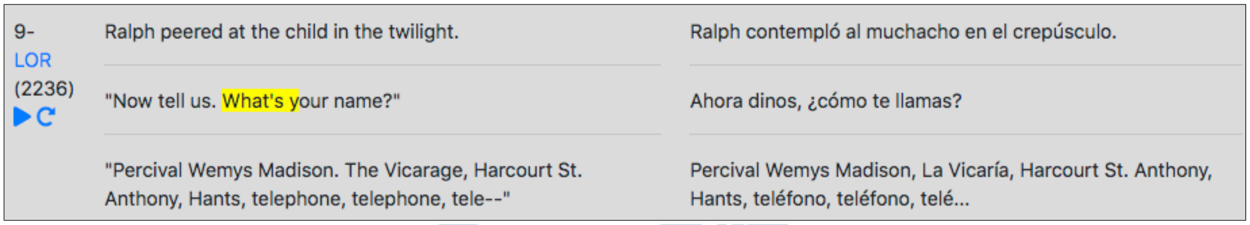
The palatization of /s/ and /z/ is a bit more complicated to examine in the corpus data as it is hard to formulate a precise query for these two phonemes by orthographical means. For example, a broad search (displaying 1 per text), such as ***s_y***, returns results from 21 different texts with mixed examples of both /s/ and /z/ in word-final position. Furthermore, only two of those 21 results show palatization. This can be dealt with two different ways.

One approach is to examine the specific examples of palatization from the broad search and carry out separate searches using those combinations. Both results from the broad ***s_y*** search showing palatization stem from the sequence *what's your*, as in *LOR (50)*, *What's your name?* and *CAT (213)*, *What's your pleasure?*. *CAT (213)* is of particular interest because it does not show /s/ and /j/ forming /ʃ/, but rather /t/ and /j/ forming /tʃ/ due to the elision of /s/ altogether. These two results can be explored further in a search for ***what's_y*** (displaying all), which yields 19 results and many more readily accessible examples of palatization than the previous broad search for ***s_y***, specifically examples showing /s/ + /j/ coalescing into /ʃ/. The first nine results come from *Lord of the Flies*, eight of which all include the same question, *What's your name?*. However, only some examples of this question show palatization while others do not. By extending the context, it is possible to see that the narrator palatizes when the character Piggy appears to ask the question, but does not do so when Ralph asks it. It seems once again that this is a deliberate choice on the part of the narrator, perhaps as a way to mark more colloquial speech. Why this choice was made could be a worthwhile topic for classroom discussion based on the characters in the book.



5- LOR (2229) ▶ ↻	Piggy knelt, holding the conch.	Piggy se arrodilló con la caracola en las manos.
	"Now then. What's your name?"	Vamos a ver, ¿cómo te llamas?
	The small boy twisted away into his tent.	El niño se fue acurrucando en su tienda de campaña. Piggy, derrotado, se volvió hacia

Figure 39. Wider context showing Piggy's dialogue in which the narrator displays palatization.



9- LOR (2236) ▶ ↻	Ralph peered at the child in the twilight.	Ralph contempló al muchacho en el crepúsculo.
	"Now tell us. What's your name?"	Ahora dinos, ¿cómo te llamas?
	"Percival Wemys Madison. The Vicarage, Harcourt St. Anthony, Hants, telephone, telephone, tele--"	Percival Wemys Madison, La Vicaría, Harcourt St. Anthony, Hants, teléfono, teléfono, telé...

Figure 40. Wider context showing Ralph's dialogue in which the narrator does not display palatization.

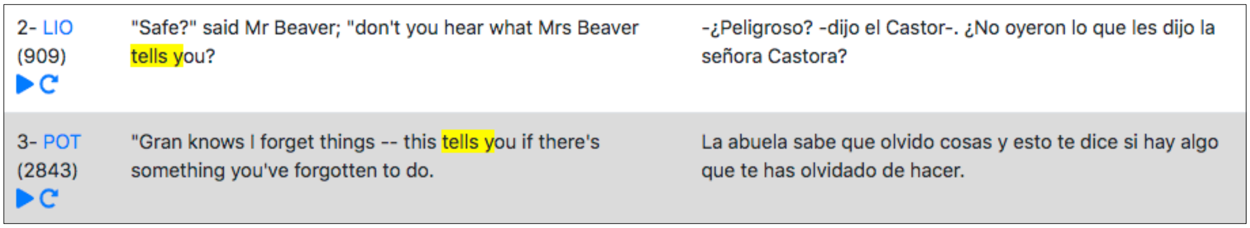
Yet, there are no other examples in the data showing complete elision of /s/ and palatization of /t/ and /j/ to /ʃ/, as occurred in *CAT (213)*.

The search for ***what's_y*** can be broadened to include all words ending in *t's* with the query ***t's_y***. This search yields 39 results, including the previous 19 from ***what's_y***, and provides three additional examples of /s/ and /j/ forming /ʃ/, *LIO (1176)*, *POT (1551)* and *POT (3075)*. In all three instances, *that's* precedes /j/. In the case of *Harry Potter*, there are two other instances of *that's* before /j/ that do not show palatization. Upon widening the context to see if this was a deliberate choice, it is unclear due to the fact that each example of *that's you* comes from a different character. This may once again be a case of the narrator's interpretation of the text, but it is difficult to say without further probing.

Another approach to locating examples of word-final /s/ and /z/, and arguably the most pedagogically effective, involves searching instances in which the penultimate phoneme in the word—that preceding -*s*-—is either voiced or voiceless as this is what will determine the phonetic realization of word-final -*s*-. This way, students not only locate more consistent examples of word-final /s/ or word-final /z/, but also gain a better understanding of assimilation, even if only word-internal. This is the reason why the previous examples from *what's* and *that's* before /j/ are examples of word-final /s/ as /s/ is preceded by the voiceless phoneme /t/ and assimilates in voicing.

Because contractions are frequent in dialogue and dialogue seems to be where most of the palatization occurs in the data, a similar search as before using contractions can be employed to locate examples of word-final /z/ before /j/. For example, a search for ***re's_y*** returns eight results, three of which show palatization of /z/ + /j/ to /ʒ/. All three examples are short sentences from spoken dialogue: *DUB (1842)*, *Where's your mother?*; *FAH (822)*, *Where's your common sense?*; *and GAT (399)*, *Here's your money*.

This same process can be carried out with non-contractions containing word-final /z/ based on assimilation. For example, a search for ***tells_y*** yields three results, all from different narrators, two of which show clear palatization.



2- LIO (909) ▶ C	"Safe?" said Mr Beaver; "don't you hear what Mrs Beaver tells you?	-¿Peligroso? -dijo el Castor-. ¿No oyeron lo que les dijo la señora Castora?
3- POT (2843) ▶ C	"Gran knows I forget things -- this tells you if there's something you've forgotten to do.	La abuela sabe que olvido cosas y esto te dice si hay algo que te has olvidado de hacer.

Figure 41. Results from the search *tells_y* in which palatization can be heard.

The same can be achieved by using only a single letter before /s/, such as -*ms*. A search for ***ms_y*** yields only one result, *HOU (2335)*, *But what is it that alarms you?*, which does, in fact, show palatization. The same can be repeated with other letters corresponding to certain phonemes before /s/.

Although searching strictly by orthography is difficult in the case of /s/ and /z/, there are two spellings that are sure to not mix the two phonemes. The first is word-final –*ss*. This spelling will only locate examples of /s/ before /j/. The query ***ss_y*** (displaying 1 per text) yields 13 results, four of which show palatization, all from dialogue. Two of the four come from the phrase *unless you*, which can then be searched separately, yielding 19 results, seven of which show palatization.

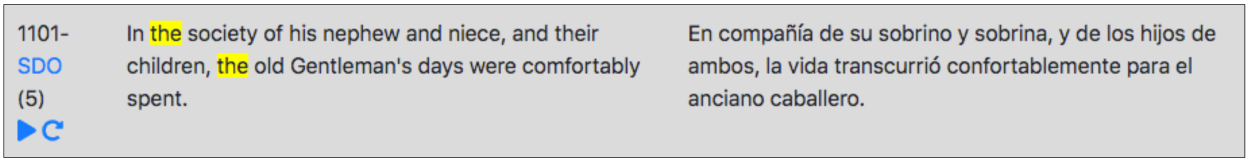
The second spelling can be found in the query ***ze_y***, which will yield instances of word-final /z/ without any interference from /s/. However, this search only returns seven results, none of which show palatization. This is possibly due to the fact that all instances involve verbs that are not as high-frequency as, say, the conjunction *unless*. Examples of the verbs found in the results include *civilize*, *authorize*, *recognize*, *criticize*, *realize*, *vaporize* and *squeeze*. Thus, it is no surprise that these verbs show no tendency toward palatization.

When it comes to palatization, it is more important for students to be aware of the phenomenon, as it may cause interference with the perception of sounds, rather than feel the need to reproduce it. The corpus data examined thus far shows that it occurs less than the alternative full realization of the two consonants next to each other. Therefore, students should not feel obligated to reproduce these examples of mutual assimilation, especially since other cases of assimilation are much more frequent and far more feasible for a Spanish speaker, as is the case with the following example dealing with the assimilation of *the*.

The definite article *the* can be realized two ways, depending on whether or not the proceeding word-initial phoneme is a vowel or a consonant. The citation form, [ðə], is normally used before consonants and semi-vowels, but changes to [ði] when occurring before a vowel. The only exception is before /i/, in which case it returns to the schwa form to avoid the same sound.

The questions proposed for the task are simple: *What are the two possible pronunciations of 'the'?* and *What determines this difference?* However, another way to propose the activity is by informing students directly of the differing pronunciations and have them use the corpus data to find out why.

Although it is possible to observe this in a broad search formulated simply as ***\bthe\b***, the problem is that there are far more instances of [ðə] due to the much higher frequency of proceeding word-initial consonants. Therefore, the teacher may want to use a specific example in which both pronunciations can be heard in the same sentence. A good example is *SDO (5)*, as seen in Figure 42 below:



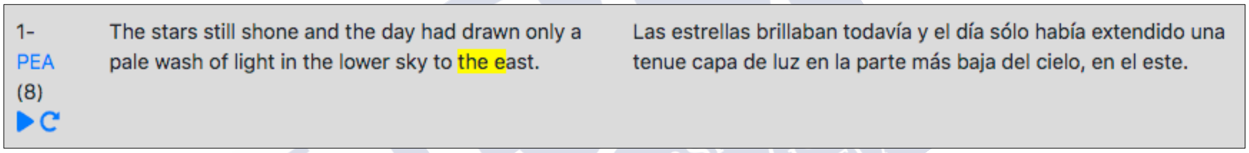
1101-
SDO
(5)
▶ ↻

In the society of his nephew and niece, and their children, the old Gentleman's days were comfortably spent.

En compañía de su sobrino y sobrina, y de los hijos de ambos, la vida transcurrió confortablemente para el anciano caballero.

Figure 42. Two instances of *the* in the same TU — one before a word-initial consonant and the other before a word-initial vowel.

Once students believe they have understood the pattern, they can confirm this tendency with the query ***\bthe_[aeiou]*** or ***\bthe_[^aeiou]***. The first result from the query ***\bthe_[aeiou]*** (displaying 1 per text) contains the example *the east*, which students can observe as the exception to the rule. They can be asked why this is, leading to further discussion and reinforcing previously acquired knowledge on this topic.



1-
PEA
(8)
▶ ↻

The stars still shone and the day had drawn only a pale wash of light in the lower sky to the east.

Las estrellas brillaban todavía y el día sólo había extendido una tenue capa de luz en la parte más baja del cielo, en el este.

Figure 43. Example of *the east* in the search results for the query *\bthe_[aeiou]*.

This concludes the present analysis of assimilation across word boundaries. The examples that have been presented here are meant to make students aware of some of the different ways this can occur as well as provide them with possible strategies for dealing with these issues in their own speech production.

The concepts covered thus far—cluster reduction, resyllabification, elision, assimilation and tone units—will be necessary for the following section on the past tense *–ed* morpheme whose phonetic expression operates under such co-articulatory influences.

3.2.1.5 The Past Tense *–ed* Morpheme

The past tense *–ed* morpheme is one of the two grammatical suffixes that marks tense in English, the other being the suffix *–s* that marks the present tense in the third person for regular verbs. Given what we have seen in terms of co-articulatory phenomena and how they can affect the realization of certain segments, it is advantageous for students to understand how the forces of connected speech may alter such a crucial suffix as one that marks the difference between past and present.

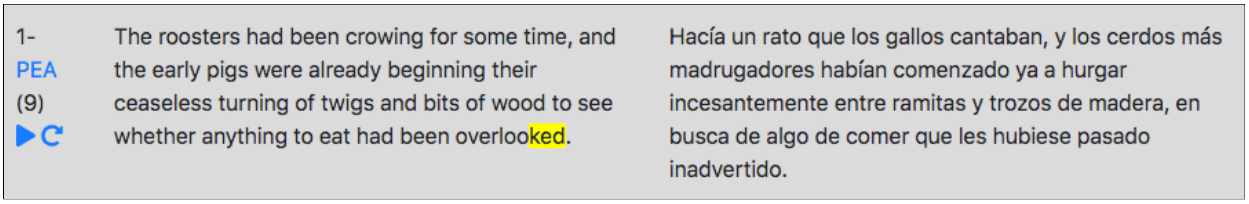
Despite being highly frequent in English and introduced very early on in language courses thanks to its simplicity in terms of formation (at least for regular verbs), the pronunciation of the past tense *–ed* morpheme often proves problematic for Spanish speakers. This is mainly due to its multiple phonetic realizations, [t, d, ɪd], and how that relates to the phonotactic constraints of Spanish.

In standard citation form, two factors determine which of the aforementioned [t, d, ɪd] will be realized: voicing and place of articulation. In the case of [t] and [d], they assimilate to the voicing of the preceding sound—[t] if it is voiceless and [d] if it is voiced, e.g. *worked* and *named*, respectively. The remaining option, /ɪd/, is realized when the preceding phoneme is /t/ or /d/ due to "the principle of separation of like sounds" leading to the "addition of an epenthetic vowel" (Pennington, 1996). Phonotactically, [d] and [ɪd] are the easiest for Spanish speakers, as long as [d] comes after a vowel, as in *played*. However, [d] preceded by another consonant is phonotactically impossible in Spanish.

On the surface this would appear to be a segmental issue since it is a morpheme whose pronunciation, as it is frequently taught to students, simply depends on the preceding phoneme within the same word. However, it will become clear that the aforementioned suprasegmental aspects will come into play when the data in the corpus is analyzed.

An examination of the past tense morpheme can be approached a number of ways. If students have never been taught the rule, they can search for all instances with the query ***ed\b***. However, this query returns an unmanageable 27,524 results whose distribution of the different realizations will likely be chaotic for students. A practical alternative to this could be to have the students search orthographically so that certain phonemes will precede *–ed*, such as /k/, /m/ or /t/, which provide examples of each possible pronunciation.

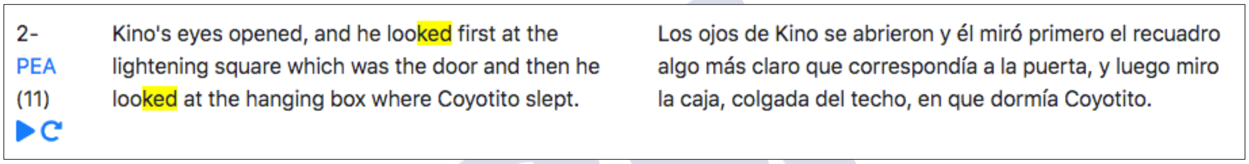
Taking a look at /k/, for example, a search for ***ked\b*** returns 3,726 results. Limiting the results to 5 per text, the first two translation units, from *The Pearl*, already provide useful information. Here we find three examples of *looked*, one in the first TU and two in the second, both from the same narrator. According to the rule, the morpheme should be realized as [t] due to assimilation to the preceding voiceless /k/. In the first TU, this can be heard quite clearly.



1- PEA (9) ▶ ↻	The roosters had been crowing for some time, and the early pigs were already beginning their ceaseless turning of twigs and bits of wood to see whether anything to eat had been overlooked.	Hacía un rato que los gallos cantaban, y los cerdos más madrugadores habían comenzado ya a hurgar incesantemente entre ramitas y trozos de madera, en busca de algo de comer que les hubiese pasado inadvertido.

Figure 44. First result from the search *ked\b*.

However, in the second TU both instances of *looked* appear to have a much more reduced pronunciation of the past tense morpheme.



2- PEA (11) ▶ ↻	Kino's eyes opened, and he looked first at the lightening square which was the door and then he looked at the hanging box where Coyotito slept.	Los ojos de Kino se abrieron y él miró primero el recuadro algo más claro que correspondía a la puerta, y luego miro la caja, colgada del techo, en que dormía Coyotito.

Figure 45. Second result from the search *ked\b*.

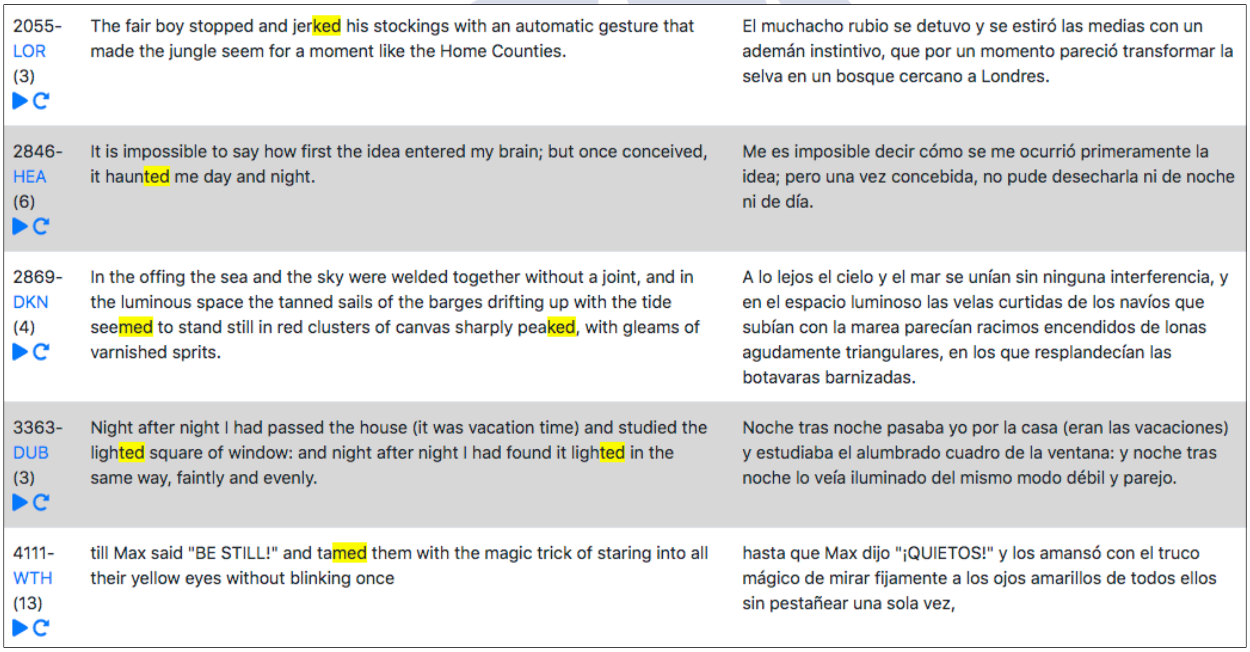
Here, the first instance of *looked* is followed by *first* and the second instance is followed by *at*. In the first TU in Figure 44, *looked* appears at the end of the sentence, which is why its full form can be heard; it is not influenced by co-articulatory phenomena as the other two examples are. This is not to say that the past tense morpheme is never fully realized when proceeded by another word, but rather the full form can be consistently heard at the end of a tone unit. To get a better idea of this, a search can be carried out for all instances of *–ked* at the end of a tone unit by placing a period, comma, question mark, or some other form of punctuation, like a colon or semicolon, after *ked*. This query can be formulated as ***ked[.,?]*** to include the most common punctuation marks, although students should be encouraged to try others if they so desire. This type of query allows the user to hear the past tense morpheme without the influence of any co-articulatory phenomena (except in those cases when a comma or semicolon is read through very quickly and there is little or no pause).

The opposite can also be done in order to examine the morpheme as it appears when followed by another word. Such a query can be formulated as ***ked_***. Keeping the results to 5 per text, the first 4 (all from *The Pearl*) provide additional information about the affects of prosody on the phonetic realization of the past tense morpheme. In the second result, *PEA (26)*, *His eyes* *flicked to a rustle beside him*, the narrator elides the past tense morpheme altogether, which would appear to be due to the fact that the same sound often gets blended into one (as seen above in section 3.2.1.2 on linking) or one segment is simply elided to facilitate pronunciation, as is the case here. The same thing happens in the fourth example, *PEA (41)*, *Kino looked down to cover his eyes from the glare*, although in this instance the following phoneme is /d/ rather than /t/. Here the elision of [t] is due to the fact that *t* and *d* share their place of articulation and pronouncing both of them, while possible, would be cumbersome and disrupt the fluidity of speech. Students should be made aware of the fact that it is possible to elide the phoneme, especially in quick, fluid speech. A common complaint among learners of English is that they do not perceive certain sounds in connected speech and this type of analysis can help them understand why certain phonemes seem to disappear. It also presents the idea of expectations. In the previous example of *Kino looked down…*, even though the past tense morpheme is elided, we know that it must be past because of the absence of the third person present tense morpheme /s/. Grammatically, it cannot be any other way and native speakers know this intuitively.

This co-articulatory phenomenon can be explored further by searching for words ending in *–ked* followed by /t/ and /d/. The query can be formulated similar to the previous one, except by adding *t* or *d* after the space (to search individually), or placing *t* and *d* in brackets (to search simultaneously). A search for only ***ked_t*** (displaying 5 per text) yields 335 results, many of which include examples of *th-* at the beginning of the proceeding words, such as *the* (examples 2, 3 & 4) and *through* (example 5). These seem to have the same kind of effect as word-initial /t/ even though the phonemes are /ð/ and /θ/, respectively. The past tense morpheme is elided in these first five results, which makes sense as the place of articulation for /θ/ and /ð/ is interdental and very near the coronal articulation of /t/ and /d/. However, because we have limited the results to 5 per text, the first five are all read by the same narrator with a North American accent. In the sixth result, *SDO (1347)*, *…other things he said too, which marked the turn of his feelings…*, a different possibility with a new narrator (now British) can be heard. The speaker fully realizes the past tense morpheme as [t] and does this repeatedly in the other translation units as well. The only instance in which the same narrator (from *Sense & Sensibility*) elides the past tense morpheme is before *to*, as in *SDO (1898)*, *…she turned away and walked to the instrument.* Students can explore this further and limit all instances to *to* via the query ***ked_to*** (displaying 5 per text). The results from *Sense & Sensibility* now appear to be much more uniform as the past tense morpheme is elided in all five instances.

It should be noted that it is essential for students to understand that these examples are not meant to be interpreted as prescriptive pronunciation models. These are tendencies, not rules. The phonological data in the corpus simply shows what speakers do and it is clear that in many cases it is not necessarily uniform. While the past tense morpheme is usually taught categorically in English language classrooms, there are a number of variables that come into play and potentially alter its realization as discussed in this section.

Rather than examining the past tense morpheme by individual searches, another option is to formulate a query that would produce all three realizations in the results, according to the rule. This can be done with the query ***[kmt]ed\b***.



2055- LOR (3) ▶ ↻	The fair boy stopped and jerked his stockings with an automatic gesture that made the jungle seem for a moment like the Home Counties.	El muchacho rubio se detuvo y se estiró las medias con un ademán instintivo, que por un momento pareció transformar la selva en un bosque cercano a Londres.
2846- HEA (6) ▶ ↻	It is impossible to say how first the idea entered my brain; but once conceived, it haunted me day and night.	Me es imposible decir cómo se me ocurrió primeramente la idea; pero una vez concebida, no pude desecharla ni de noche ni de día.
2869- DKN (4) ▶ ↻	In the offing the sea and the sky were welded together without a joint, and in the luminous space the tanned sails of the barges drifting up with the tide seemed to stand still in red clusters of canvas sharply peaked, with gleams of varnished sprits.	A lo lejos el cielo y el mar se unían sin ninguna interferencia, y en el espacio luminoso las velas curtidas de los navíos que subían con la marea parecían racimos encendidos de lonas agudamente triangulares, en los que resplandecían las botavaras barnizadas.
3363- DUB (3) ▶ ↻	Night after night I had passed the house (it was vacation time) and studied the lighted square of window: and night after night I had found it lighted in the same way, faintly and evenly.	Noche tras noche pasaba yo por la casa (eran las vacaciones) y estudiaba el alumbrado cuadro de la ventana: y noche tras noche lo veía iluminado del mismo modo débil y parejo.
4111- WTH (13) ▶ ↻	till Max said "BE STILL!" and tamed them with the magic trick of staring into all their yellow eyes without blinking once	hasta que Max dijo "¡QUIETOS!" y los amansó con el truco mágico de mirar fijamente a los ojos amarillos de todos ellos sin pestañear una sola vez,

Figure 46. Results from the query *[kmt]ed\b*.

This query utilizes brackets to include all examples in which /k/, /m/ and /t/ (along with /ð/ and /θ/ in the cases of *th-*) are the preceding phonemes. Students can then be asked what might be influencing the realizations of the past tense morpheme. To facilitate perception and avoid interference from the proceeding segments as seen above, students may be better off searching for examples at the end of tone units. Such a search could be formulated as ***[kmt]ed[.,?]***. In doing so, data from the examples with *–ked* and *–ted* clearly confirm the rule, pronounced as [-kt] and [-tɪd] respectively, while data from *–med* is not so clear. The first five instances (displaying 5 per text) of *–med* come from 4 different narrators and a slight aspiration can be heard in each case, appearing more voiceless than voiced. A search for ***med[.,?]*** would confirm this tendency at the end of a tone unit, complicating what appeared on the surface to be a simple rule that depended on voicing.

Therefore, rather than looking to the end of tone units for fully realized instances of the past tense morpheme, another option is to examine what happens when the proceeding word begins with a vowel, as this will likely lead to resyllabification of the morpheme into a syllable- and word-initial position. Such a search can be formulated as ***med_[aeiou]*** (displaying 1 per text). The data shows that now the past tense morpheme is almost invariably realized as [d]. Only one narrator realizes what sounds to be more like [t] than [d] due to aspiration, but this is likely the result of pauses she makes before continuing, therefore creating a tone boundary. This can be heard in the first result from *DKN (199)*, *She <u>seemed uncanny</u> and fateful*. The narrator makes a significant pause which allows for the following word to have little to no effect on the pronunciation of the past tense morpheme, much like the examples at the end of a tone unit.

Returning to the original example of *–ked*, it is possible to see how all proceeding consonants will affect the past tense morpheme with the search query ***ked_[^aeiouwy]*** (displaying 1 per text), which returns all instances of letters that are *not* vowels or semi-vowels. There is little consistency in these examples in terms of full realization and elision, except for when the following word is a pronoun that begins with *h*, such as *him* or *her*. If observed by students or pointed out to them, this can shed light on the elision of *h* in word-initial position as well as the reduction of function words to their weak forms more generally. In *SDO (173)*, *…and she <u>liked him</u> for it*, the word-initial /h/ is elided leading the word-final /t/ from the past-tense morpheme to become resyllabified and the aspiration to come through clearly. The general lack of consistency in the data is still informative for students nevertheless as it forces them to reevaluate the cues they listen for in connected speech.

Examining how the past tense morpheme is affected by prosodic features is perhaps most useful for comprehension. This way, students know not to rely on only a single cue, i.e. the clear pronunciation of the past tense morpheme, but rather consider other factors such as grammatical conditions and prosodic speech. Students can see that in many cases, particularly those read with more caution, the narrator fully realizes the pronunciation of the morpheme. Nevertheless, such an analysis may also prove beneficial for their own speech production as they can locate models in the data that provide strategies for dealing with difficult co-articulation, as seen in the previous section on assimilation across word boundaries. This also includes clusters within words as well, such as *asked*.

Although not a matter of prosody per se, the previous searches would expose students to certain consonant clusters resulting from the addition of the *–ed* morpheme. One of the most ubiquitous examples of this is the complex VCCC cluster in *asked*. The cluster /skt/ is phonotactically impossible in Spanish, and therefore extremely difficult to pronounce for Spanish speakers (along with many other learners of English). Because *asked* is so frequent in English literature, the corpus is full of examples, many of which appeared in translation units from the above searches, as in ***ked[.,?]***. A simple query formulated as ***\basked\b*** yields 692 results (displaying 1 per text). A simple task can be formed from this: students can be asked if it is possible to reduce or simplify this cluster in any way. The data in the corpus shows that it can indeed be simplified by eliding the middle consonant, resulting in [æst], as in *LIO (148)*, *But a moment later she <u>asked</u>, Mr. Tumnus....* It is precisely these types of simplification strategies on which the data can inform the student.

All of the cases that have been analyzed in this section can be turned into DDL exercise questions, similar to the training exercises from the video tutorials. Teacher intervention is essential to guiding students in the right direction. Success also depends on how much prior knowledge students have regarding English phonology and the features of connected speech. Therefore, some of the questions suggested below may not be applicable. Furthermore, a broad question such as question 1 may require additional information depending on the students' level of English, e.g. "There are four possibilities".

1. Set the Results option to 1 per text. Carry out a search with the query *ed\b* in English. In what ways can the regular past tense ending *–ed* be pronounced?
2. Carry out a search with the query *[kmt]ed\b*. How does the preceding sound effect the pronunciation of *–ed*?

3. Set the Results option to 5 per text. Carry out a search with the query *ked\b*. Is there a difference when *–ed* appears at the end of a sentence and when it is followed by another word?
4. Is there a difference in the way *–ed* is pronounced if the following words begins with a consonant or a vowel?
5. Is there any way to simplify the pronunciation of *asked*?

Other classroom activities can be designed to explore these nuances in connected speech. One possibility is a dictation exercise using a TU in which the past tense morpheme is elided. Another is to have the students locate examples for themselves in the corpus, such as when *-ed* is realized as [t]. How the corpus is exploited will depend heavily on factors such as students' comfort with the corpus, the ease with which they can search and locate data, their level of English, their experience with the language, whether or not they have been taught the past tense morpheme before, etc. This is why there is no single way in which the questions should be asked. Teachers must tailor the questions and activities to each individual learning context.

This type of approach to the past tense morpheme is meant to raise students' awareness about how English prosody can alter segments to the point where sounds change or are elided in order to maintain fluid speech. The examples in which the most careful speech can be heard are usually those that have the least amount of alteration in the realization of the past tense morpheme.

This type of activity embodies the true spirit of DDL, as what began as a simple investigation into the pronunciation of the past tense morpheme quickly turned into a much deeper and nuanced look at suprasegmental features. It is easy for a controlled, hands-on DDL exercise to morph into a more exploratory, hands-off activity. Exploring the corpus data inevitably leads to unexpected questions that shed new light on topics students may have been unaware of before. In this instance, by drawing their attention to the past tense morpheme, they inevitably encounter other, broader aspects of connected English speech that influence the realization of certain phonemes.

3.2.1.6 Vowel Reduction and Weak Forms in Connected Speech

Vowel reduction is what gives English its characteristic rhythm. A reduction in vowel quality often means a reduction in stress, therefore giving prominence to other, usually stressed, and therefore fuller, vowels. This is another reason why certain words, particularly function (or grammatical) words, seem to disappear in rapid English speech. Spanish does not have this tendency. This can make connected English speech difficult to follow, and even more difficult to imitate and reproduce in a native-like way. This section will examine how vowel reduction can be explored in the corpus data and in what ways that information is useful to learners of English. To do so, instances of the modal verb *can* will be examined in the corpus data, followed by a look at other modal verbs such as *have to*, and *should*, which will give way to an analysis of the cliticized *a*.

First, it is necessary for students to be able to distinguish between function words and content words, as it is the former that frequently undergo vowel reduction in order to maintain isochrony. As stated above, function words are those words which are not nouns, verbs, adjectives or adverbs, such as prepositions, articles, pronouns, etc. Vowel quality is inextricably linked to the differentiation of function and content words. This is because function words have two forms, a citation form and the reduced, or weak form. In fact, Field (2003) identifies fifty-one words with weak forms, most of them function words. Field (2008) also notes that as a result, "learners face an important obstacle in distinguishing content and functors when their L1 does not resemble English rhythmically" and that "[t]here is thus a conflict between high-frequency/familiarity and low perceptual evidence". This puts Spanish-speakers at a disadvantage. Vowel quality, particularly when it comes to unstressed words, can have major effects on intelligibility (Solé Sabater, 1991), as will be seen in the following analysis of *can*.

In English, the "default" stress setting is for content words to be stressed and function words to be unstressed and therefore undergo reduction. It makes sense for content words to receive greater stress as they generally carry the semantic weight and are more pertinent to the message of the utterance. This pattern can change, of course, if the speaker wants to show contrast or emphasis, which means that any word of a sentence has the potential to be stressed (Solé Sabater, ibid), depending on what the utterance is attempting to convey.

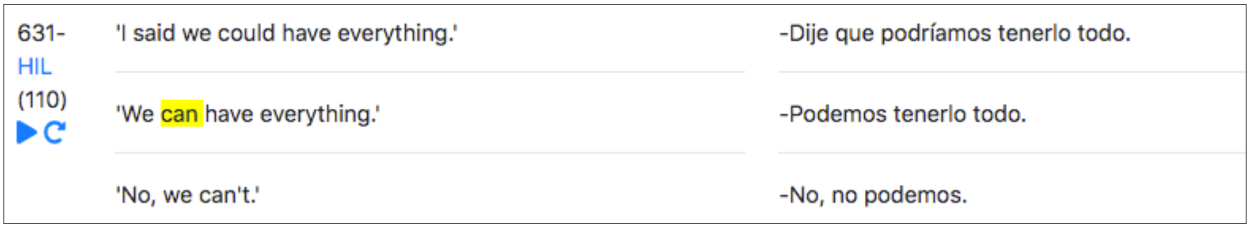
Besides stressed vowels having longer duration and being generally louder in order to make them stand out, vowel reduction of unstressed syllables is a third element that contributes to prominence (Solé Sabater, ibid). This is achieved by reduction to a middle vowel, usually /ə/ or /ɪ/ (Field, 2003), two vowels that do not exist in Spanish. Therefore, it is important for students to learn that many of the vowels in an utterance need not be fully realized as would normally be the case in their citation forms.

The modal verb *can* is usually learned very early on due to its high-frequency and overall utility. Students are most likely familiar with the full forms of *can*, /kæn/ (NAmEng), /kan/ (RP). However, students are less likely to be familiar with the reduced forms, /kən/ or /kɪn/, unless they have been taught these explicitly. Furthermore, the reduction of *can* allows for the negative contraction *can't* to be differentiated more easily as it is never reduced.

An examination of the weak forms of *can* in the corpus data may be approached different ways. Students may be asked the broad question *What are the two pronunciations of 'can' according to the corpus?*, although the differences may not be noticeable enough for many students who have never been exposed to the idea of vowel reduction. Therefore, it may be more advantageous to first discuss the weak and full forms and then ask *When is one form used and when is the other?*. Rather than carry out general searches and hope the students figure it out on their own, it may be more beneficial to guide them through the searches, pointing out specific instances and comparing and contrasting them. For example, a search for ***\bcan\b[^']***[36] (displaying 1 per text) returns a variety of instances perfect for analyzing how prosody affects the pronunciation of *can*, as it can be heard before other words, at the end of a tone unit, and stressed for emphasis. These different instances produce different pronunciations. The majority of results show *can* in the middle of the sentence, which in most cases produces the weak form as can be heard in the first four examples. However, in the fifth example, *LOR (175)*, *Soon as he can*, *can* appears at the end of the sentence, and therefore at the end of a tone unit and can be heard in its full form. It is also of longer duration than the previous four weak forms. Another instance of *can* at the end of a tone unit occurs in *HUN (26)*, *If you can*. Here, the same full form can be heard. Of special note are those results in which *can* is heard in its full form but not at the end of a tone

[36] Because apostrophes, or single quotes, are used for quotations to mark dialogue, they are not recognized as forming part of a word. Therefore, a search for ***\bcan\b*** will return all instances of *can't* as well.

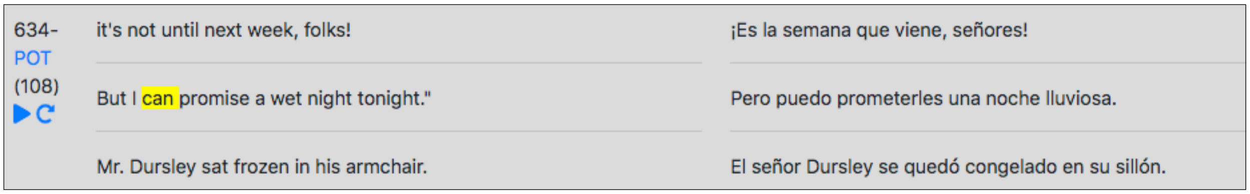
unit. This occurs in *HIL (110)*, *We can have everything*, and *POT (108)*, *But I can promise a wet night tonight*. By widening the context, one can get a better understanding of why *can* is not reduced. With the ability to see how these sentences operate in context, it can be concluded that *can* is stressed in both cases to show contrast. In *HIL (110)*, the full form shows contrast to what the other person says in the dialogue, that they *can't*, as can be seen in Figure 47.



631- HIL (110) ▶ ↻	'I said we could have everything.'	-Dije que podríamos tenerlo todo.
	'We can have everything.'	-Podemos tenerlo todo.
	'No, we can't.'	-No, no podemos.

Figure 47. First example in which *can* is stressed to show contrast.

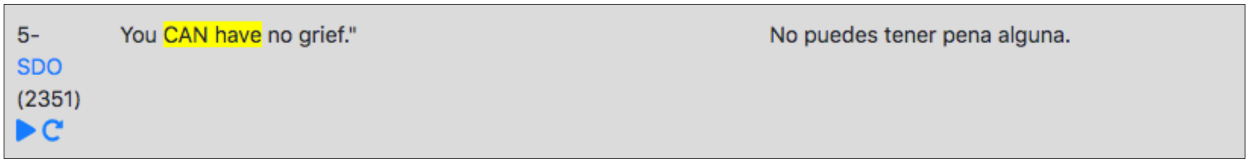
In *POT (108)*, the contrast is made between *next week* and *tonight* and what they are able to do, as seen in Figure 48.



634- POT (108) ▶ ↻	it's not until next week, folks!	¡Es la semana que viene, señores!
	But I can promise a wet night tonight."	Pero puedo prometerles una noche lluviosa.
	Mr. Dursley sat frozen in his armchair.	El señor Dursley se quedó congelado en su sillón.

Figure 48. Second example in which *can* is stressed to show contrast.

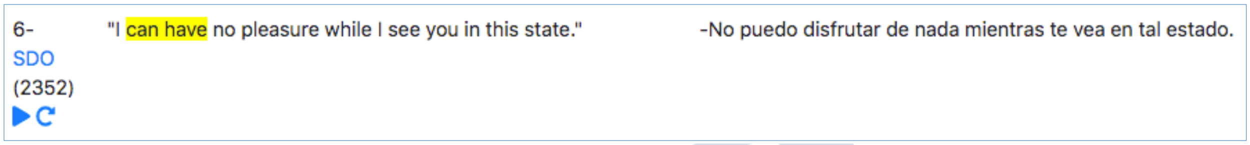
Because *can have* is likely a high-frequency item, at least when compared to *can promise* (a quick search comparison shows this difference: 20:2, respectively), students can carry out a follow-up search. A search for ***\bcan_have\b*** yields 20 results, including the example from figure 47 above. While nearly all the others reduce *can* and place the focus on *have*, another example can be located in which not only is *can* not reduced, but it is even written to reflect its stress, as seen in Figure 49 below.



5- SDO (2351) | You CAN have no grief." | No puedes tener pena alguna.

Figure 49. Result showing *can* written to show emphasis.

Furthermore, the following TU in the search results is also the very next TU within the same text, as seen in Figure 50.



6- SDO (2352) | "I can have no pleasure while I see you in this state." | -No puedo disfrutar de nada mientras te vea en tal estado.

Figure 50. The TU following *SDO (2351)* shown in Figure 49.

The user can compare the two instances from the same conversation within the text. The narrator follows the cue in SDO (2351) and pronounces the full form of *can*, while in the following TU it is reduced as expected.

Therefore, while *can* is normally reduced to its weak form through vowel reduction and shortening in connected speech, students can see that it is possible for it to be stressed, and also what such stress implies, both semantically and phonetically.

Further searches can be carried out to find all instances at the end of a tone unit, in which case the full form is consistently produced. This is because the weak form of *can* is anticipatory, giving prominence to the main verb to follow. However, when there is no main verb to follow, there is no need for reduction. The query ***\bcan[.,?!;]*** yields 64 results and shows that the full form of *can* at the end of an tone unit is categorical. This type of search also reinforces the idea of tone unit boundaries and what effect they have on vowel realization. For example, in *SDO (238)*, *no one <u>can</u>, I think, be in doubt*, the clause *I think* separates *can* from the main verb, but because it forms its own tone unit, there is a boundary after *can*, and the full form helps the listener perceive it.

Another option is to search for the noun form of *can*. This can be accomplished two ways using both versions of LITTERA: the CLUVI version and the SensoGal version. To locate the noun form in the CLUVI version, the query can be formulated as ***\ba_can\b|the_can?\b|\bcans\b***.

This locates both the singular and plural forms and yields 8 results, all showing the full vowel. However, this search will not find any instances of *can* in the singular preceded by an adjective, such as *a red can*. Additional regular expressions can be used to locate instances of *can* as a noun preceded by a non-article. The query can be formulated as ***\ban?_\w*_can\b***. However, as the results demonstrate, there is nothing to prevent *\w** from being a noun and *can* continuing to function as a modal verb. Of the 15 results for ***\ban?_\w*_can\b***, only two noun forms of *can* are located, both from the same text and the same noun phrase, *trash can*. All the other results show *can* functioning as modal before noun phrases such as *a boat*, *a dream*, *a body*, etc.

The other approach is to search SensoGal using the Part of Speech option. Searching for the noun *can* by lemma, however, only yields seven results, one less than the original query involving regular expressions in the CLUVI version. The missing TU is OLD (1666), *a can of coffee*. After locating this TU from another search, it seems *can,* in this instance, has been incorrectly tagged as a verb. While the exercise in SensoGal is still effective, this example serves as reminder as to the inherent limitations of relying on automatic tagging when operating within a pedagogical context.

Finally, students can compare the weak form of *can* in regular connected speech to the consistent full-form pronunciation of *can't* in the corpus data. This distinction between *can* and *can't* is a common issue with learners of English, especially Spanish speakers whose first language does not use weak and strong forms to achieve a more or less even isochrony. Learners who are unaware of the idea of weak and full forms might then ask how it is possible to tell the difference between the two. What complicates the matter even further is the fact that the /t/ in *can't* often goes unreleased or elided altogether, thus appearing to vanish in the sequence of sounds. After examining the corpus data, it becomes apparent that it is the full vowel that clues in the listener, and therefore, a full form in mid-sentence is most likely the negative. A full form at the end of a sentence or at an intonation boundary could be either the negative or affirmative, although meaning is then discerned by context, e.g. *No, you can't*, or the presence of a word-final /t/, possibly aspirated. The idea is to make learners aware of the same cues that native-speakers pick up on naturally.

The data shows similar findings for *have* when used as a modal. Depending on the speed with which one speaks, *have* can have different reduced forms, usually involving an elision of the initial /h/, such as /əv/, /əf/ or even simply /ə/. Because *have* also functions as a main verb and not always as an auxiliary, it is easy to compare instances of full and weak forms in a single search with the query ***\bhave\b*** (displaying 1 per text). This comparison can even be made within the same TU in some instances, as occurs in *LOR (22)*, *Some of them must have got out. They must have, mustn't they?*. Both the weak and full vowel form can be heard within this TU, despite the fact that /h/ is elided in both instances. *Have* as an auxiliary preceding a participle is reduced to /əv/, but then appears in its full form at the end of a tone unit before the comma, just like in the previous examples with *can*. Another instance of *have* at the end of a tone unit is *HIL (20)*, *No, you wouldn't have*, in which the full form can be heard.

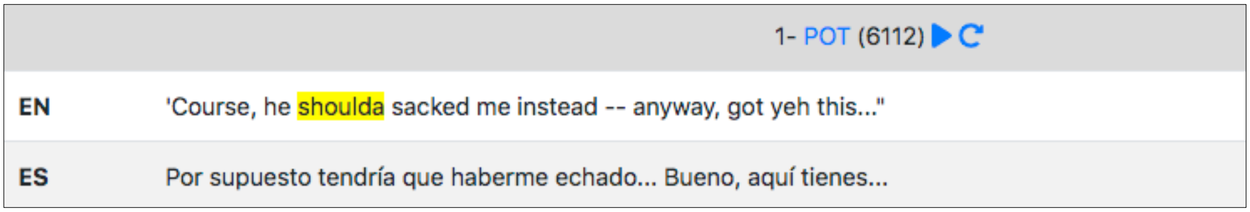
However, there are instances in which *have* functions as an auxiliary verb but shows no reduction, as in *FAU (13)*, *…where the two boards would have met*. This is possibly due to the fact that the audio samples are scripted speech and some narrators make attempts to be as clear and articulate as possible, to the point where their speech may even sound "too clean" to provide adequate models for conversational English. Of course, it is not prescriptively incorrect to realize *have* in its full form when functioning as an auxiliary, but such models may not adequately prepare learners for the realities of fluid, connected speech. Nevertheless, having both weak and full forms present in the data allows students to make comparisons. Unlike *can*, using the full form of *have* will rarely lead to it being mistaken for the negative *haven't* due to the fact that *haven't* contains two syllables and /v/ is resyllabified, becoming the onset of the next syllable.

A good example of vowel reduction and linking though resyllabification in the data is *FAH (41)*, *I'd have known it with my eyes shut*. In this case, /d/ is resyllabified as the word-initial /h/ is elided. The vowel is also reduced, resulting in [aɪ.dəv.noʊ.nɪt].

Whether or not auxiliary *have* will be realized as [əv] or [əf] depends on the proceeding sound due to assimilation. *NIN (117)* provides an example of both in the same TU, *to have lost* and *to have forgotten*. In the former, *have* is reduced to /əv/, as it is proceeded by a voiced /l/. In the latter, *have* is realized as /əf/ due to the voiceless /f/ in word-initial position following *have*. This is why it is important to establish the principles of linking, elision and assimilation early on as so many phenomena in connected English speech are best explained through these concepts.

Before moving on, it is necessary to briefly introduce the idea of cliticization. Cliticization in phonology occurs when an unstressed word attaches itself to a neighboring stressed word for the sake of rhythm (Whitley, 2002; Field, 2003). Field provides the example of *want to* realized in speech as *wanna*. Clitics can also be found in contractions, e.g. *should've*, as will be seen below. Cliticization will be an important aspect of the following analysis.

To find more examples of *have* as an auxiliary, students can search for instances in which it follows modals such as *should*, *would*, *could* or *must*. This can be done individually or with a batch search. For instance, a search (displaying 1 per text) for examples of *should have* followed by a participle can be formulated as ***\bshould_have_\w*e[dn]\b***. The data shows that in cases of rapid speech, it is possible for *have* to be reduced to schwa, as in *POT (2531)*, *…should have seen their faces…*, [ʃʊ.də.si:n]. The reduced form can also be compared to the contraction, *should've*, which yields six results, all from the same two books, *Harry Potter* and *The Fault in our Stars*[37]. Naturally, the contraction leads to reduced forms, [əv] and [əf], as to be expected when *have* is functioning as a clitic. The authors of the texts were certainly aware of this and use it to create the effect of natural, connected speech in dialogue. A follow-up question may then be: *Is it possible for 'have' to be reduced only to [ə] in should've?. If so, how might this be represented orthographically?*. This is accomplished by adding *a* to the end of the word, in this case *shoulda*, hence the colloquial expression *shoulda-woulda-coulda* in English to express the idea of not dwelling on the past. A search for *shoulda* yields a single result, again from *Harry Potter*. Besides showing the reduction to /ə/, the same TU also shows another way in which the author attempts to reflect colloquial English orthographically by spelling *you* as *yeh*.



1- POT (6112)

EN	'Course, he shoulda sacked me instead -- anyway, got yeh this..."
ES	Por supuesto tendría que haberme echado... Bueno, aquí tienes...

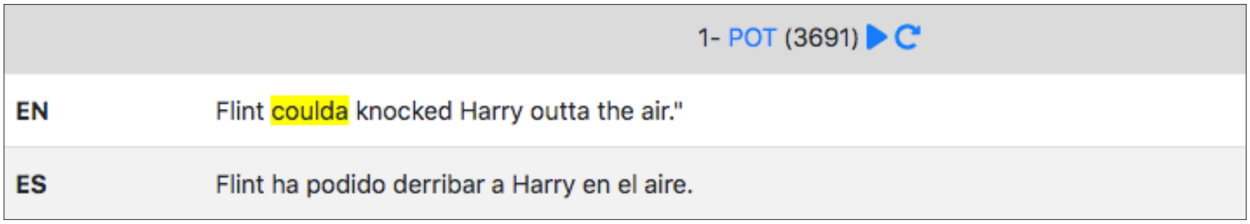
Figure 51. ***Should have*** **reduced to** ***shoulda*****. Also note** ***you*** **spelled as** ***yeh*** **to reflect colloquial speech.**

If students were to separately search *yeh* in the corpus, it would yield 99 results, although all from *Harry Potter*. What makes this example even more illuminating is the fact that there is no explicit subject in the sentence, *Anyway, got yeh this*. This deletion of the first person subject is a

[37] This was the advantage of including both of these works in the corpus. They contain more colloquial registers and orthography to reflect them.

possibility in informal English speech, yet goes against everything students are taught regarding the importance of explicit subjects (except in commands) and the rigidity of these rules. This demonstrates the value of authentic language, which makes DDL so effective as it raises students' awareness of such aspects of real usage that are rarely discussed in language classrooms and not included in fabricated textbook examples.

The clitic *a* can be examined alongside other modals, such as *could* and *would*, although only the former yields a result, as seen in Figure 52.



1- POT (3691)

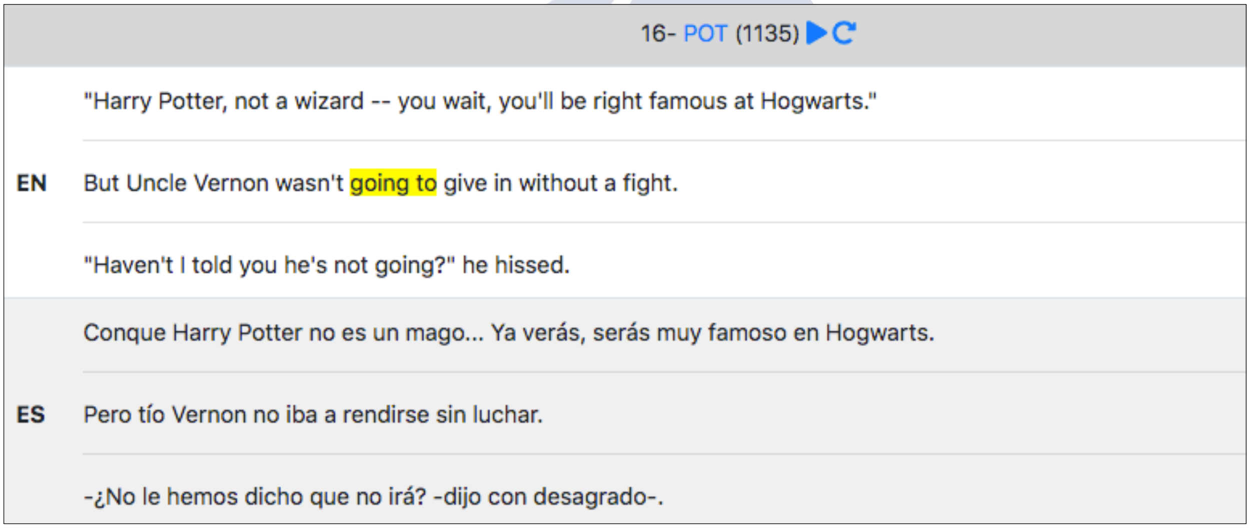
EN	Flint coulda knocked Harry outta the air."
ES	Flint ha podido derribar a Harry en el aire.

Figure 52. The lone search result for *coulda*.

Much like in the previous example with *shoulda*, here we can find additional information on English speech. In this case, there is another example of cliticization, although now as a result of *out of* being reduced to *outta*. Not only is *have* reduced to a cliticized schwa, so is *of*. Once again, *Harry Potter* is the only book containing this orthographic representation of fluid English. Despite being scripted speech, the narrators attempt to follow the cues provided by the author's deliberate spelling. Seeing two different examples in which words are reduced to schwa and cliticized in the same TU provides learners with valuable insights into informal speech.

It is also worth pointing out the Spanish translation of the cliticized forms. Because there is no equivalent of such reduction in Spanish, the translations for *should have*, *should've* and *shoulda* all capture the grammatical sense of the original text (using *deber* + *haber*, e.g. *Debí haberte dicho* for *I should've told you*), but make no attempt to capture the colloquial nature of *should've* and especially *shoulda*.

Other important examples of clitics in the form of schwa can be explored the corpus, such as *gonna*, *gotta* and *wanna* for *going to*, *got to* and *want to*. In these instances, the word that is reduced to schwa is not *have* or *of*, but *to*. These, along with other examples of cliticization, are regular features of conversational English achieved through vowel reduction. To search for both *gonna* and *wanna* simultaneously, the query can be formulated as ***\b[gw][oa]nna\b***. All the results come from two texts, *Harry Potter* and *The Fault in Our Stars*, and are pronounced as would be expected, [gʌ.nə] and [wa.nə]. The results can then be compared to those from a search for the full forms from the same narrators by searching in each text individually and using the query ***going_to|want_to***. In *Harry Potter*, the full form can be heard, although, interestingly, in some instances the narrator pronounces *going to* as *gonna*, as in *POT (1135)*, *But Uncle Vernon wasn't going to give in without a fight*. And an instance of *wanna* can be heard in *FAU (450)*, *But I want to show Hazel Grace….* In the case of *FAU (450)*, the cliticization is in a line of dialogue, which is not surprising. However, in *POT (1135)*, the narrator is speaking between two lines of dialogue as seen below.



16- POT (1135) ▶ ↻

EN

"Harry Potter, not a wizard -- you wait, you'll be right famous at Hogwarts."

But Uncle Vernon wasn't going to give in without a fight.

"Haven't I told you he's not going?" he hissed.

ES

Conque Harry Potter no es un mago... Ya verás, serás muy famoso en Hogwarts.

Pero tío Vernon no iba a rendirse sin luchar.

-¿No le hemos dicho que no irá? -dijo con desagrado-.

Figure 53. The context surrounding *going to* when realized as *gonna*.

The line is read quickly, perhaps not to lose the momentum of the back and forth in the dialogue. Rapid speech is likely what led to the reduction.

It would appear that when it comes to examples of colloquial speech in *Harry Potter*, "birds of a feather flock together", as many of the results for *gonna* contain other representations of colloquial speech, such as double negatives, *'cause* instead of *because*, *don'* in place of *don't*, etc. Again, none of this comes through in the translations. For more on representations of informal speech in the corpus, see the following section 3.2.1.7 below.

As for *gotta*, a simple query by spelling reveals that it not only appears in *Harry Potter* and *The Fault in our Stars*, but in *Fahrenheit 451* as well in *FAH (250)*, *we gotta go*, and occurs in dialogue, as would be expected. The same pronunciation of [ga.ɾə][38] can be heard in all the results. The full form can be heard via the query ***\bgot_to\b***, in which the reduced form is sprinkled throughout by different narrators, as in *LOR (4501)*, *We got to get out of this*, and *FAH (655)*, *Well, I've got to be going*. These, like many examples of cliticization, occur in dialogue and are most likely the result of the narrator attempting to provide "realistic" or "authentic" portrayals of English speech.

This section has aimed to show how important it is for students to understand the concept of vowel reduction in English and how real instances can be examined and contrasted in the LITTERA corpus. While they may not necessarily need to reproduce them, students should be aware of the fact that certain words, mainly those with grammatical functions, have two forms, a weak form and a full form. The weak form is created through vowel reduction, either to /ə/ or /ɪ/, which often leads to the cliticization of certain words, such as the auxiliary verb *have* or the prepositions *to* and *of*. This aspect of English speech is indispensible for oral comprehension.

3.2.1.7 Representations of Speech in Dialogue

When a spoken corpus or a speech corpus is created and the audio is transcribed, the orthographic transcription may reflect some of the aforementioned features of natural English speech, such as palatization (e.g. *got you* written as *gotcha*) and cliticization (e.g. *going to* written as *gonna*). Similarly, LITTERA is composed of fictional texts, many of which contain dialogue written to reflect the way people speak, at least from the authors' perspectives as native speakers of the language. Some examples of this have already been mentioned in the previous section, such as the cliticization of *should have* to *shoulda*. When speech is represented orthographically in a literary text, the audiobook narrator is prompted to use the pronunciation that reflects the written dialogue, resulting in the kinds of phenomena characteristic of connected English speech. This section will look at four different ways authors in the corpus represent English speech in

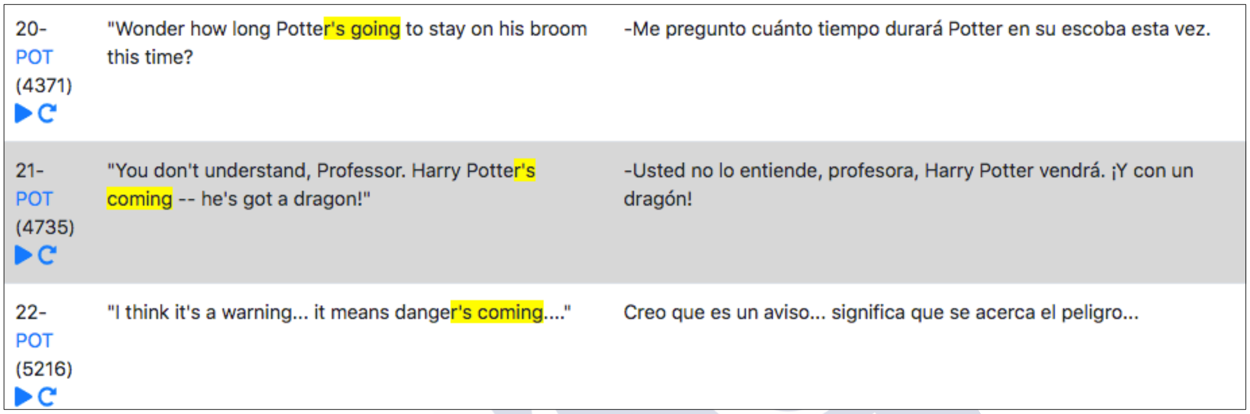
[38] Note the alveolar tap (commonly known as a *flap*) in the place of /t/. This occurs frequently in North American English when /t/ appears intervocalically. In the case of *gotta*, however, there is no difference between varieties of English. See section 3.2.1.8 below for more on flaps.

dialogue: through 1) contractions, 2) additional examples of the cliticized *a*, 3) the elision of word-initial and word-final letters and 4) alternative spellings to indicate vowel reduction.

A good way to see how speech is represented in dialogue is through the use of contractions. When used in written dialogue, contractions aim to reflect the conversational pronunciation of English. Students are generally introduced to contractions early on when they learn how to form the commonly used contraction *it’s* or negative sentences with *don’t*, *doesn’t* and *didn’t*. Yet, there are many ways English speakers regularly use contractions in connected speech that students may be less aware of.

The verb *to be* is frequently contracted in connected speech. Students learn many of its most common contractions early on, such as *I’m*, *it’s*, *you’re*, etc. However, students should be made aware of how it can combine with other words, such as *wh-* question words or when nouns, both common and proper, are the subjects of the verb. Because a very common way to form such contractions is with *is*, the task can be prompted with the question *Can ‘is’ form contractions with wh- words?* or *Can ‘is’ form contractions with nouns?*. To find contractions with *wh-* words, a batch search can be carried out using the query **_\bwh\w*’s\b_** (displaying 5 per text). It becomes readily apparent from the results that the most common forms are *what’s*, *where’s* and *who’s*. While the pronunciation of *what’s* and *where’s* is fairly clear, an issue arises with *who’s*, as it can misinterpreted as its homophone *whose*. These two can be examined more closely in a separate search formulated as **_who’s|whose_**. If students are unfamiliar with *whose*, its meaning can be deduced from the translation as *cuyo* along with its derived forms *cuya*, *cuyos* and *cuyas*. The pronunciation does not differ between *who’s* and *whose* in the data, making the grammatical context essential to understanding the meaning of the sentence. The corpus data shows that *whose* is always followed by a noun, while the two visible examples of *who’s* are followed by the present participle *writing* and the adjective *afraid*.

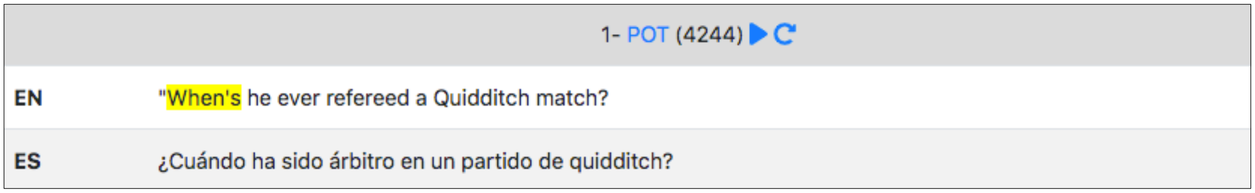
Returning to the search results for contracted *wh-* words, there is an instance in which the contraction is not formed from a question word, but rather a common noun. This occurs in *GAT (873)*, *But the wheel’s off*. This tells students that it is possible, especially in conversational English, to combine *is* with a noun. Unfortunately, this is not an easy phenomenon to explore on its own in the corpus due to the fact that searching by **‘s** will yield an overwhelming number of *‘s* marking the possessive. That’s not to say that individual searches, such as ***r’s*** will not provide any results, but rather the results for contractions will be buried among those showing the possessive. One way to eliminate some of the unwanted possessives is to search for *r’s* before the present participle, as in ***r\’s_\w*ing***. Although many possessives still show up, the three results from *Harry Potter* all show the contraction of *is* with a noun, both proper and common, as seen in Figure 54 below:



20- POT (4371) ▶ ↻	"Wonder how long Potter's going to stay on his broom this time?	-Me pregunto cuánto tiempo durará Potter en su escoba esta vez.
21- POT (4735) ▶ ↻	"You don't understand, Professor. Harry Potter's coming -- he's got a dragon!"	-Usted no lo entiende, profesora, Harry Potter vendrá. ¡Y con un dragón!
22- POT (5216) ▶ ↻	"I think it's a warning... it means danger's coming...."	Creo que es un aviso... significa que se acerca el peligro...

Figure 54. Examples of contractions formed with nouns and *to be*.

The two forms that do not immediately appear in the results for contractions with *wh-* question words are *why* and *when*. The original batch search can be recalibrated, of course, to show all the results instead of 5 per text, but a more efficient way is to search them directly. A search for *why’s* yields only three results and a search for *when’s* yields only one. However, the result for *when’s* sheds light on another issue; that is, not only can *‘s* reflect the possessive and *to be* contractions, it can also occur from the third person singular *have* functioning as an auxiliary verb. Figure 55 shows the lone search result for *when’s*:

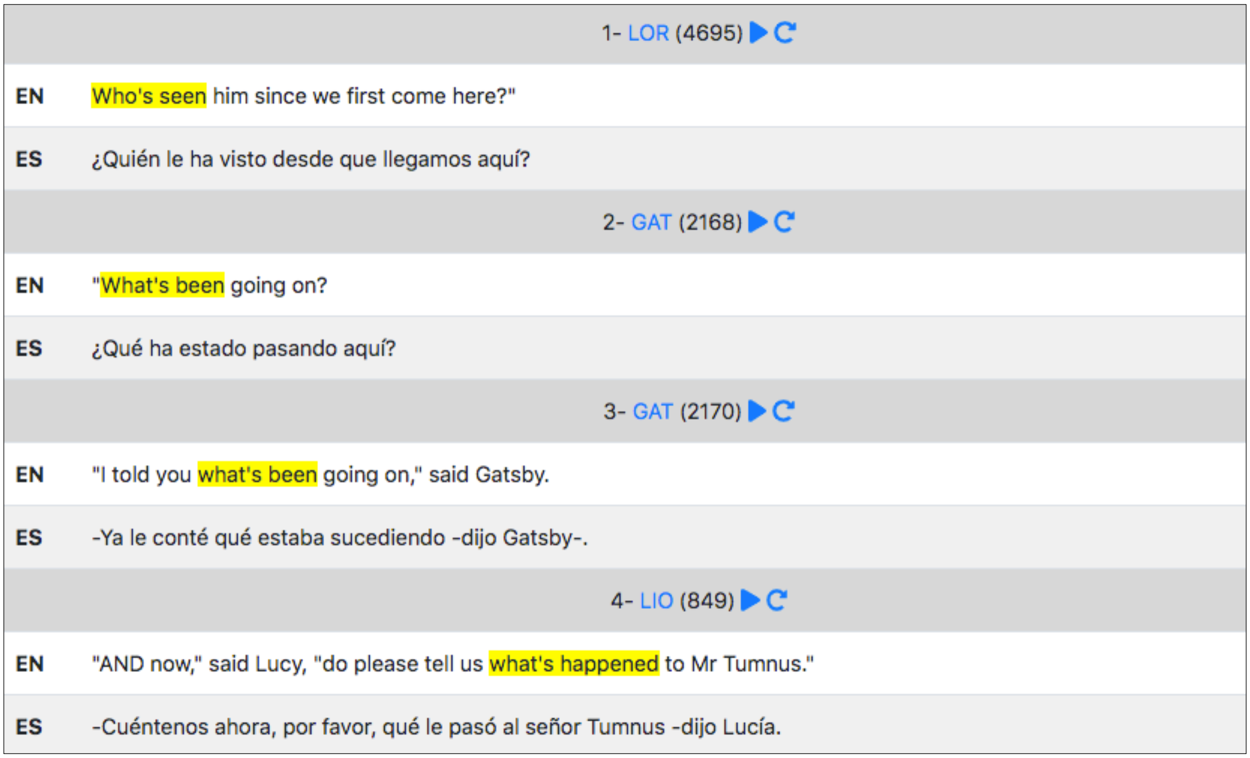


1- POT (4244) ▶ ↻	
EN	"When's he ever refereed a Quidditch match?
ES	¿Cuándo ha sido árbitro en un partido de quidditch?

Figure 55. Search result for *when’s* showing *when* forming a contraction with *has*.

This also occurred in the batch search for *to be* contracted with *wh-* questions, as in *FAH (785), Who’s got a match!*. The contraction is part of the verb phrase *has got*. The same occurs in *GAT*

(918), *What's that got to do with it?*. To explore this further in the corpus, students can utilize the Spanish search function and accompany the original English query of ***\bwh\w*'s\b*** with ***\bha\b***. The search yields eight results, two of which still show contractions with *to be*. Another option is to add the past participle to the search query, as in ***\bwh\w*'s_\w*e[dn]\b***. The string *\w*e[dn]\b* is used to find any word that ends with *–ed* or *–en*, which are the endings for most past participles. This query also returns eight results, only one of which is not a contraction of *have*, but rather the passive.



1- LOR (4695) ▶ ↻

EN Who's seen him since we first come here?"

ES ¿Quién le ha visto desde que llegamos aquí?

2- GAT (2168) ▶ ↻

EN "What's been going on?

ES ¿Qué ha estado pasando aquí?

3- GAT (2170) ▶ ↻

EN "I told you what's been going on," said Gatsby.

ES -Ya le conté qué estaba sucediendo -dijo Gatsby-.

4- LIO (849) ▶ ↻

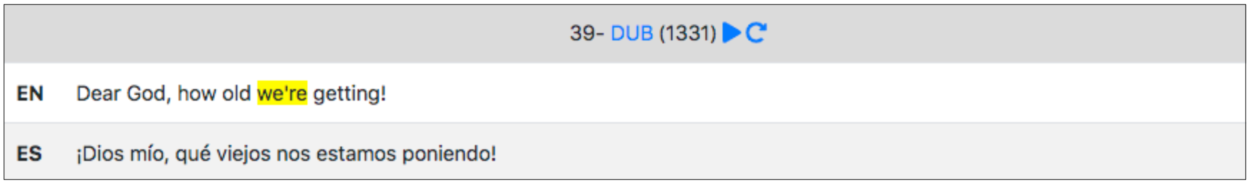
EN "AND now," said Lucy, "do please tell us what's happened to Mr Tumnus."

ES -Cuéntenos ahora, por favor, qué le pasó al señor Tumnus -dijo Lucía.

Figure 56. First four search results for the contracted form of *have* in the corpus.

One final way to find contracted forms of *has* is by eliminating the *wh-* question word and using personal pronouns for the third person singular, such as *he*, *she* and *it*. The query can be formulated as ***\bit's_\w*e[dn]\b|he's_\w*e[dn]\b***. While the search query may appear complex, it is simply building upon previous queries. Also, note that no boundary was placed before *he's*. This allows for *she's* to appear in the results as well. This search yields 62 results, and although instances of *is* + *adjective* can be seen, there are also many more examples of contractions with *has* than in the previous two attempts using *wh-* words. *Been* appears frequently in the results, especially when proceeding *it's*. Students unfamiliar with this contraction may mistake its pronunciation for *its bin** or *its pin** in English speech. It is essential for students to be familiar with these forms as they are common in spoken English. A possible follow-up exercise that can be drawn from these examples of *has* reduced to its contracted form is a dictation in which the students must then translate the sentence, therefore having to decide if the contraction is a form of *is* or *has*.

What may also cause confusion when dealing with contractions of *to be* is that their pronunciation in fluid speech may be difficult to perceive or simply mistaken for another word, as in the case of *we're* due to the fact that [wiə] or [wiɚ] can be reduced to [wə] or [wɚ] and sound similar or identical to *were*. This can be heard in the corpus with the query ***we're*** (displaying 1 per text). In *DUB (1331)*, *we're* is pronounced identical to *were*.



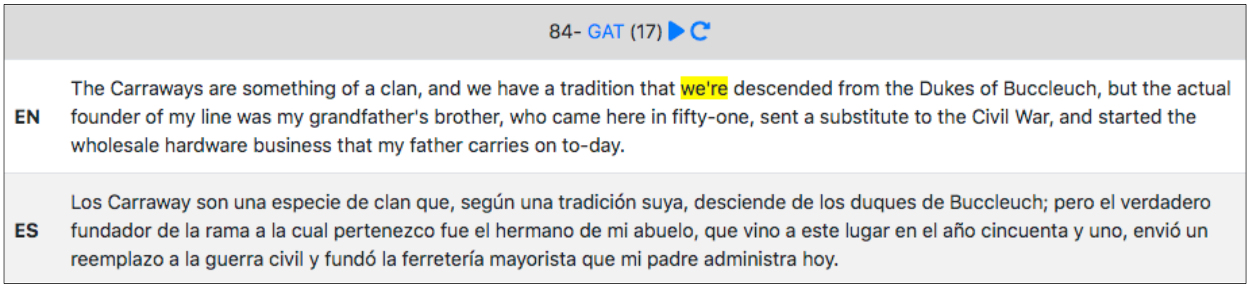
39- DUB (1331)

EN Dear God, how old we're getting!

ES ¡Dios mío, qué viejos nos estamos poniendo!

Figure 57. Search result for *we're* in which *we're* is pronounced identically to *were*.

Of course, *were* is grammatically impossible here and native speakers will intuitively perceive this as *we're*. However, two results down, in *GAT (17)*, *we're* appears after the relative pronoun *that*, which may lead to confusion in terms of subject-verb agreement due to the fact that *tradition* is singular and the learner may interpret *we're* to be *were*.



84- GAT (17)

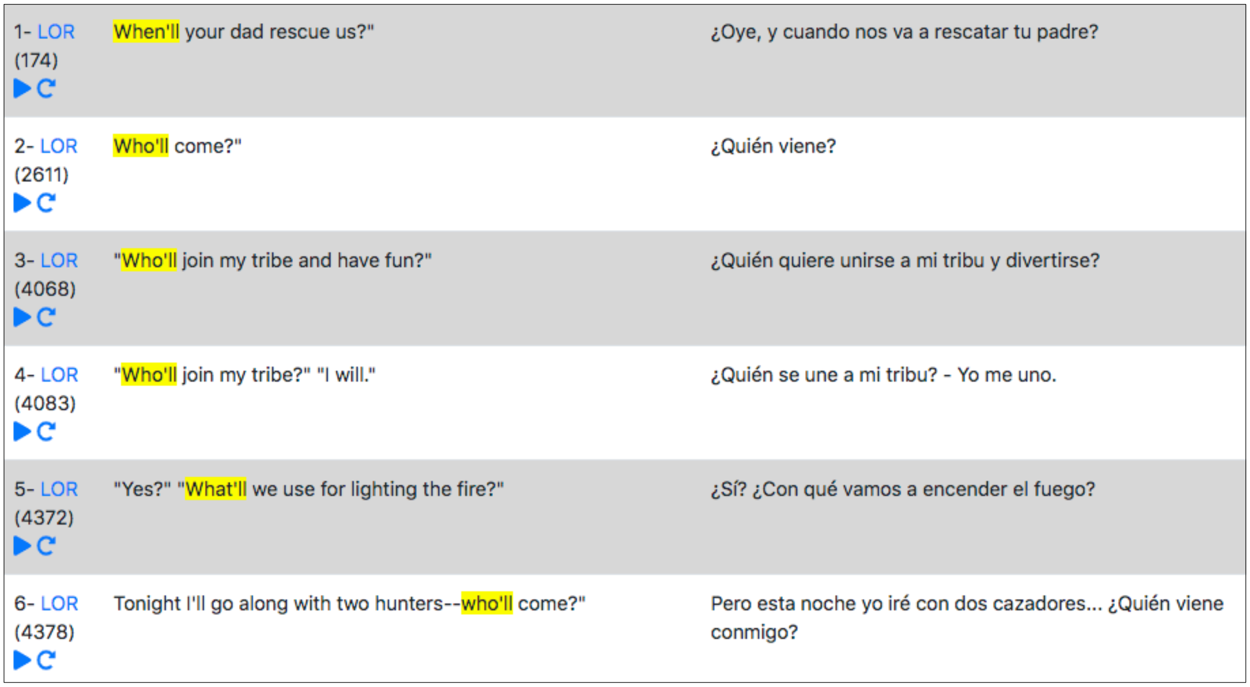
EN The Carraways are something of a clan, and we have a tradition that we're descended from the Dukes of Buccleuch, but the actual founder of my line was my grandfather's brother, who came here in fifty-one, sent a substitute to the Civil War, and started the wholesale hardware business that my father carries on to-day.

ES Los Carraway son una especie de clan que, según una tradición suya, desciende de los duques de Buccleuch; pero el verdadero fundador de la rama a la cual pertenezco fue el hermano de mi abuelo, que vino a este lugar en el año cincuenta y uno, envió un reemplazo a la guerra civil y fundó la ferretería mayorista que mi padre administra hoy.

Figure 58. Example in which *we're* appears after a relative pronoun and may cause confusion in agreement.

This can become a DDL activity through the prompt *What are the two possible pronunciations of ‘we’re’? Find examples of each.* Another way to approach the exercise might be to ask *What word might ‘we’re’ be confused with? Provide examples from the corpus data.*

Once again, the speech cues students assume will guide them in listening, e.g. the full form of *we’re*, may be (and frequently are) different from their citation form in connected speech. This becomes especially evident with contractions involving the modal verbs *will* and *would*. Contractions with these modal verbs are fairly easy to locate in the corpus. To find examples of *will*, the query can be formulated as ***‘ll\b*** (displaying 5 per text). This search returns a total of 838 results. The data show that most of the contractions are formed with personal pronouns, which is not surprising. This is how students often learn the future with *will*, by practicing examples from nouns and personal pronouns. However, three other words appear in the results to form contractions with *will*: *what* (*what’ll*), *there* (*there’ll*) and *that* (*that’ll*). As students are less likely to be familiarized with these forms, separate searches can be carried out to see and hear more examples of each. A search for *what’ll* yields 6 results. It is good for students to be exposed to cases such as this one because it is a clear example of the modal *will* being reduced to a dark /l/, [ɫ], leading to [wʌ.ɾəɫ] or [wʌɾ.ɫ], in which /l/ is a syllabic consonant. The /t/ is flapped in all instances and therefore sounds like /d/. This may inspire a broader search involving *wh-* question words, much like in the previous examples above. This search query may be formulated as ***\bwh\w*’ll\b*** and yields 11 results, including *when’ll* and *who’ll* along with *what’ll*, all exhibiting the same dark /l/ in place of *will*.



1- LOR (174) ▶ ↻	When'll your dad rescue us?"	¿Oye, y cuando nos va a rescatar tu padre?
2- LOR (2611) ▶ ↻	Who'll come?"	¿Quién viene?
3- LOR (4068) ▶ ↻	"Who'll join my tribe and have fun?"	¿Quién quiere unirse a mi tribu y divertirse?
4- LOR (4083) ▶ ↻	"Who'll join my tribe?" "I will."	¿Quién se une a mi tribu? - Yo me uno.
5- LOR (4372) ▶ ↻	"Yes?" "What'll we use for lighting the fire?"	¿Sí? ¿Con qué vamos a encender el fuego?
6- LOR (4378) ▶ ↻	Tonight I'll go along with two hunters--who'll come?"	Pero esta noche yo iré con dos cazadores... ¿Quién viene conmigo?

Figure 59. Search results showing contractions with *wh-* question words and *will*.

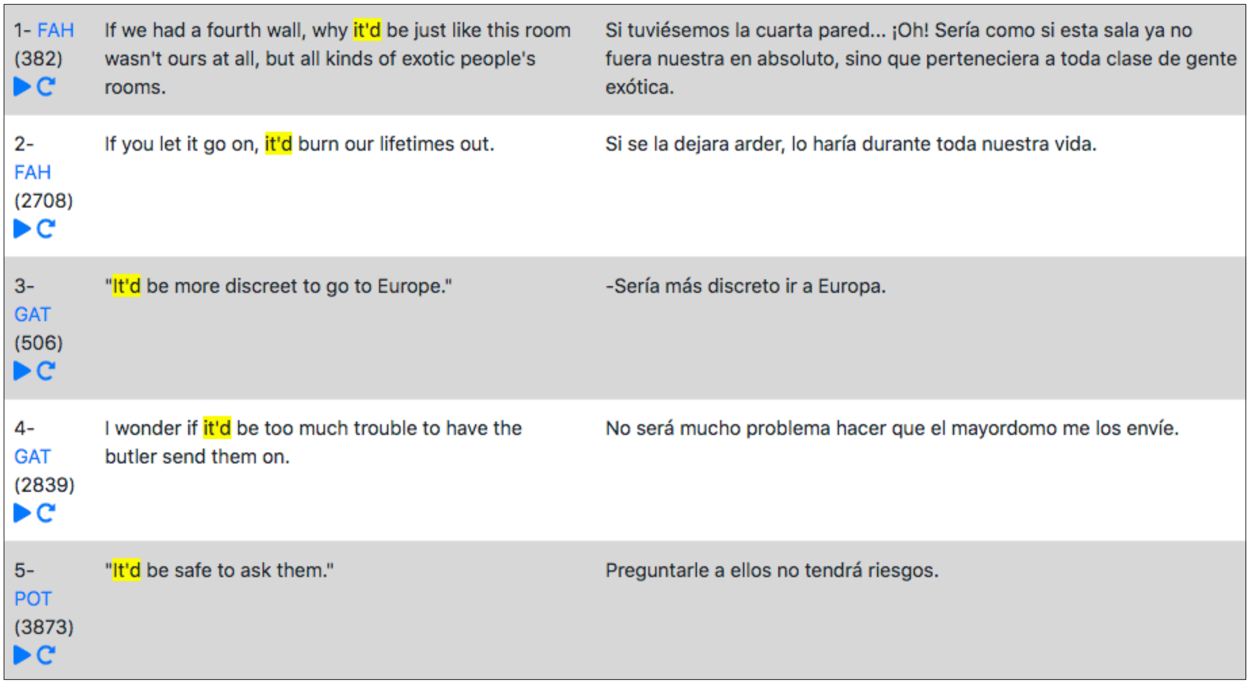
In the case of *there'll*, a search returns seven results from four different texts. In some instances, /r/ seems to be elided altogether, resulting in a pronunciation similar to *they'll*, as in *LOR (4931), Now you'll eat and there'll be no smoke.*

A search for *that'll* yields 13 results from six different texts. Similar to *what'll*, the /t/ before the apostrophe is flapped in most cases, although in 3 examples, /t/ can be heard coming through. In one such example, *POT (4864), That'll take a lot of explaining, that* is stressed for emphasis and realized as [ðæ.təɫ]. Nevertheless, the same narrator pronounces it with a flap on three other occasions. Interestingly, the other two instances in which *that* is not flapped occur when *that* is functioning as a demonstrative pronoun and not a relative pronoun. This is most likely due to the fact that *that* as a relative pronoun is often reduced in connected speech while *that* as a demonstrative is never reduced.

Returning to the contractions of *will* with personal pronouns, many of these require little examination beyond the frequency with which they are used as students are certainly familiar with them. Yet, there are some instances in which students may struggle to both perceive and produce the contraction, such as the third person singular forms *it'll*, *he'll* and *she'll*. This is due to unexpected pronunciations. Searching *it'll* in the corpus data (displaying 1 per text), the same flapping can be heard as in *what'll*, sounding similar to the North American pronunciation of *little*. In fact, *it'll* can be compared with *little* by carrying out a simultaneous search for both in the corpus, limiting the results to American English. Students who are unaware of this reduction will have a difficult time understanding the contraction. Being able to recognize and reproduce [ɪ.ɾəɫ] is advantageous for students, especially those aiming to speak in a more "native-like" way. Flapping will be discussed in further detail in the following section, 3.1.2.8.

Both *he'll* and *she'll* can be explored simultaneously in the same search with the query ***he'll\b*** (displaying 5 per text). By leaving the beginning open and not specifying a word boundary, all cases of *she'll* will be included in the results. Both a full form and a reduced form of each can be heard. In the full form, *he'll* and *she'll* are realized as [hi:ɫ] and [ʃi:ɫ], respectively. An example of the full form of *he'll* can be heard in *LOR (36)*, *He'll be back all right*, and a full form of *she'll* can be heard in *GAT (1773)*, *She'll see*. They can be compared to the words *heel* and *shield* in terms of vowel quality. In the reduced form, however, the vowel is reduced to [ɪ], leading to [hɪɫ] and [ʃɪɫ], homophonous to the words *hill* and *shill*. These can be heard in *FAH (1614)*, *He'll come in*, and *DUB (1508)*, *She'll have a good fat account…*.

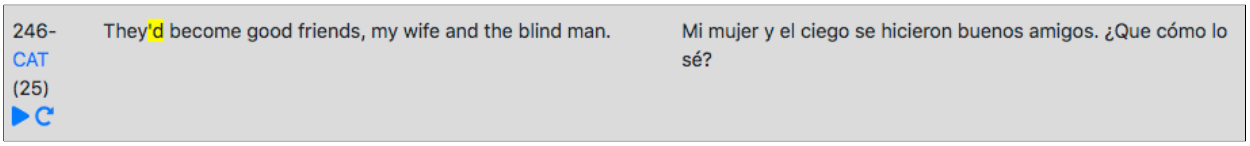
The other modal discussed here, *would*, can be examined in the same way as *will*. A broad search for ***'d\b*** (displaying 5 per text) returns 783 results, showing students that both *will* and *would* are frequently contracted in connected speech. In terms of pronunciation, contractions formed with *would* present only minor difficultly to students as it is merely adding /d/ to the end of the word. If the word ends in a vowel, as in almost all personal pronouns, this is not an issue. If it ends in a consonant, this can potentially cause problems, such as with the personal pronoun *it*, and the contraction *it'd*. Because /t/ and /d/ are both coronal consonants, it is phonotactically impossible in English to pronounce *it'd* as it is written. When searched in the corpus, the data from the five results for ***it'd*** shows that it can be pronounced two ways. Firstly, an epenthetic vowel can be added so that both /t/ and /d/ are realized, that is [ɪ.təd], just like the past tense *–ed* morpheme when added to a word ending in /t/. This can be heard in the first three results. Alternatively, the /t/ can be elided altogether to form [ɪd], as heard in the final two results.



1- FAH (382) ▶ ↻	If we had a fourth wall, why it'd be just like this room wasn't ours at all, but all kinds of exotic people's rooms.	Si tuviésemos la cuarta pared... ¡Oh! Sería como si esta sala ya no fuera nuestra en absoluto, sino que perteneciera a toda clase de gente exótica.
2- FAH (2708) ▶ ↻	If you let it go on, it'd burn our lifetimes out.	Si se la dejara arder, lo haría durante toda nuestra vida.
3- GAT (506) ▶ ↻	"It'd be more discreet to go to Europe."	-Sería más discreto ir a Europa.
4- GAT (2839) ▶ ↻	I wonder if it'd be too much trouble to have the butler send them on.	No será mucho problema hacer que el mayordomo me los envíe.
5- POT (3873) ▶ ↻	"It'd be safe to ask them."	Preguntarle a ellos no tendrá riesgos.

Figure 60. First five search results for *it'd*.

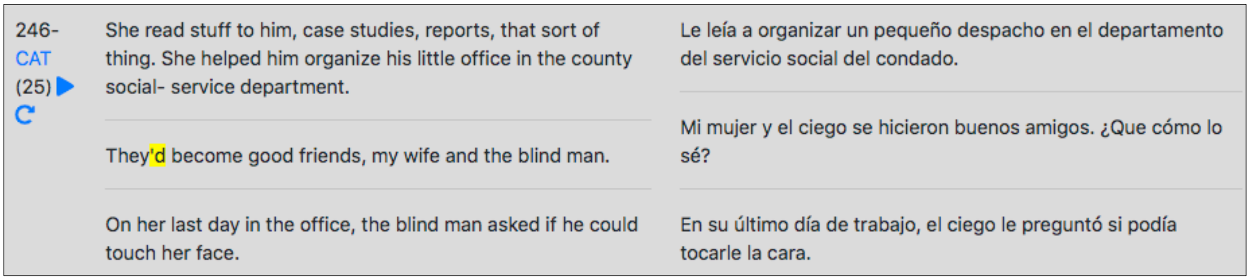
Students must be careful when examining the results from the broad search ***'d\b*** if the intention is to find contracted forms of *would*, as it is possible to also find contracted forms of the auxiliary *have* in the past tense, *had*. Normally, it is not difficult to differentiate between the two grammatically, as *would* is followed by an infinitive and *had* is followed by the past participle. However, because there are certain verbs which are the same in both the infinitive and participle form (e.g. *come*, *become*, *run*, etc.) it is a good idea to make students aware of these possibilities. An example of this can be found in *CAT (25)*:



246- CAT (25) ▶ ↻	They'd become good friends, my wife and the blind man.	Mi mujer y el ciego se hicieron buenos amigos. ¿Que cómo lo sé?

Figure 61. Example of a contraction with *had* which could be mistaken for *would*.

Judging by the translation, it is safe to assume that it is a contraction with *had*. This can be cross-checked by widening the context, as in figure 62 below:

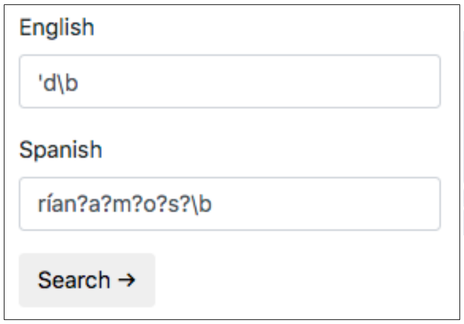


246- CAT (25) ▶ ↻	She read stuff to him, case studies, reports, that sort of thing. She helped him organize his little office in the county social- service department.	Le leía a organizar un pequeño despacho en el departamento del servicio social del condado.
	They'd become good friends, my wife and the blind man.	Mi mujer y el ciego se hicieron buenos amigos. ¿Que cómo lo sé?
	On her last day in the office, the blind man asked if he could touch her face.	En su último día de trabajo, el ciego le preguntó si podía tocarle la cara.

Figure 62. Example of how the context can be used to disambiguate the contraction *they'd*.

Students should be made aware of these potentially confusing situations in spoken English, and, thanks to the dialogue in the corpus, this is possible.

To locate only instances of *would* in contractions and avoid *had*, the Spanish search option can be employed. One way to formulate the query could be the following:



English

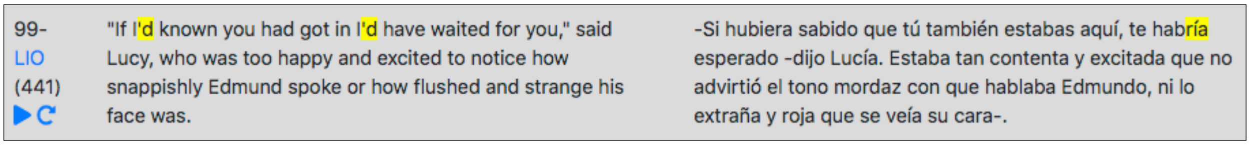
'd\b

Spanish

rían?a?m?o?s?\b

Figure 63. Bilingual search query to only return examples corresponding to *would*.

However, this query will still allow certain instances of contracted *had* to slip through, as in *LIO (441)*:

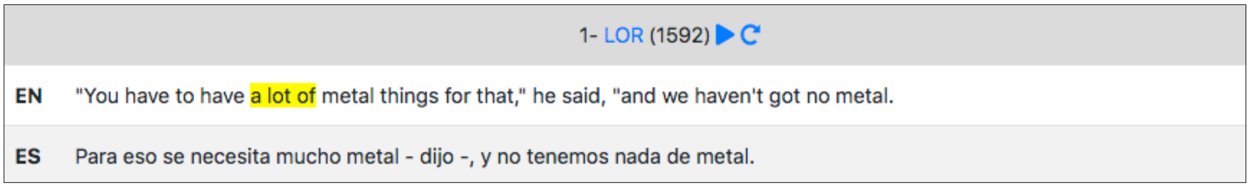


99- LIO (441) ▶ ↻	"If I'd known you had got in I'd have waited for you," said Lucy, who was too happy and excited to notice how snappishly Edmund spoke or how flushed and strange his face was.	-Si hubiera sabido que tú también estabas aquí, te habría esperado -dijo Lucía. Estaba tan contenta y excitada que no advirtió el tono mordaz con que hablaba Edmundo, ni lo extraña y roja que se veía su cara-.

Figure 64. Example of contracted *had* still turning up in the search results.

In this example, the past perfect is followed by the present perfect conditional in the English text. However, the highlighting in the Spanish translation guides the user toward the correct equivalent.

There are still a few other representations of English speech in written dialogue to be analyzed here. As seen in the previous section, *outta* is a way to pronounce *out of* in connected speech by reducing *of* to schwa and making it a clitic. Other examples that were provided include *gonna*, *wanna*, *shoulda* and *gotta*. If students want to explore this further and find instances in which *of* or *to* are cliticized in a similar way, a query can be formulated to return results with *–tta*, such as ***tta\b***. However, because the first 25 results are from the name *Gretta*, the search can be adjusted accordingly to ***[^e]tta\b***. This query returns 34 results, many of which have already been discussed here, such as *gotta* and *outta*. Two new results appear as well, *atta* and *a lotta*. *Atta* is found in *POT (422)* in the expression *Atta boy*, which is a reduced form of *that a boy* and is read by the narrator as written, [æ.tə.boi]. This is a colloquial expression, translated adequately as *bravo*, and is not found in its full form within the corpus. *A lotta* is a cliticized form of *a lot of*. There is only one instance of *a lotta* in the results, *POT (1607)*, and it is read accordingly with a flap, [ə.la.ɾə]. When *a lot of* is searched in the corpus (displaying 1 per text), the same pronunciation can be heard from the dialogue in the first result, *LOR (1592)*:



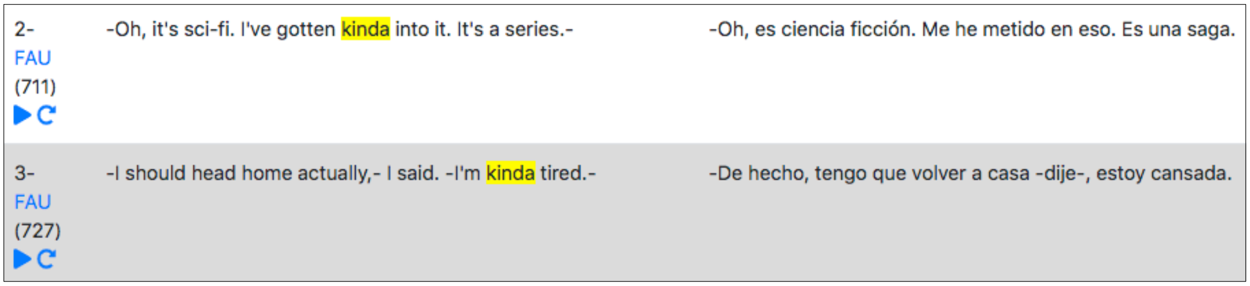
1- LOR (1592) ▶ ↻

EN "You have to have a lot of metal things for that," he said, "and we haven't got no metal.

ES Para eso se necesita mucho metal - dijo -, y no tenemos nada de metal.

Figure 65. Example of *a lot of* in which the narrator pronounces the cliticized form *a lotta*.

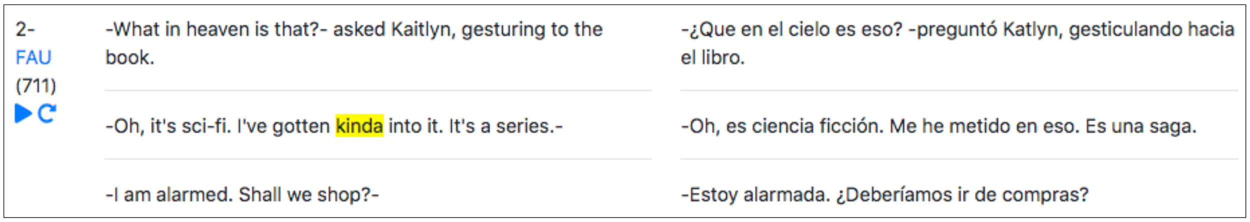
There are two more examples of cliticization in the written dialogue that can be found in the corpus data and have not been discussed. These are *kinda* and *sorta*. Both contain a reduction of the preposition *of*, similar to *out of* from above. A search for *kinda* yields 7 results from two books, *Harry Potter* and *The Fault in Our Stars*. Because six of the seven are from the latter, a search for *kind of* within the same book can allow the learner to hear both forms, as the narrator is consistent in her pronunciation of one and the other following the cues of the written dialogue. Additionally, an analysis of the translations for the seven results for *kinda* sheds some light on how *kinda* can be used in discourse. For instance, the second and third results have no equivalent in the translations, as seen below in Figure 66.



2- FAU (711) ▶ ↻	-Oh, it's sci-fi. I've gotten kinda into it. It's a series.-	-Oh, es ciencia ficción. Me he metido en eso. Es una saga.
3- FAU (727) ▶ ↻	-I should head home actually,- I said. -I'm kinda tired.-	-De hecho, tengo que volver a casa -dije-, estoy cansada.

Figure 66. Results for *kinda* with no equivalents in the translations.

The second and third, said in dialogue, are most likely the result of hedging. By extending the context, this becomes evident.

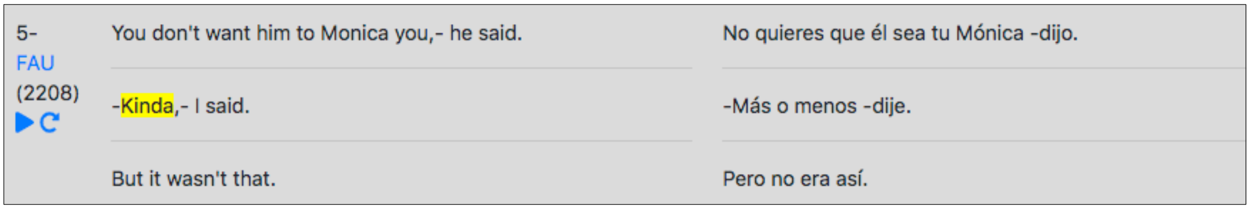


2- FAU (711) ▶ ↻	-What in heaven is that?- asked Kaitlyn, gesturing to the book.	-¿Que en el cielo es eso? -preguntó Katlyn, gesticulando hacia el libro.
	-Oh, it's sci-fi. I've gotten kinda into it. It's a series.-	-Oh, es ciencia ficción. Me he metido en eso. Es una saga.
	-I am alarmed. Shall we shop?-	-Estoy alarmada. ¿Deberíamos ir de compras?

Figure 67. Extended dialogue showing hedging.

It seems *kinda* is used to downplay the speaker's new interest in science fiction, most likely due to embarrassment given the other person's reaction of shock.

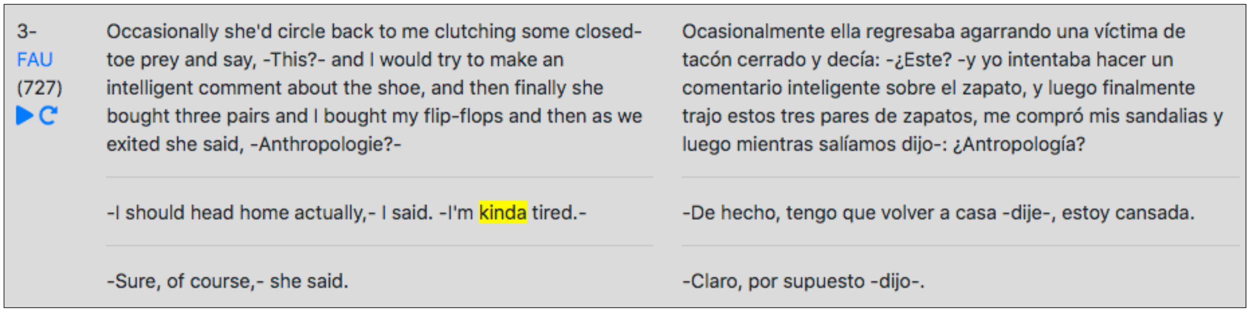
Yet, in other instances, such as *FAU (2208)*, the context shows us that it can be a way to indirectly say no, possibly out of politeness.



5- FAU (2208) ▶ ↻	You don't want him to Monica you,- he said.	No quieres que él sea tu Mónica -dijo.
	-Kinda,- I said.	-Más o menos -dije.
	But it wasn't that.	Pero no era así.

Figure 68. Example of *kinda* as a way to show politeness.

The first-person narrator admits (*But it wasn't that*) that what she said to the other person in the dialogue was not true, supporting the idea that *kinda* was used as a polite way to not say *yes*, something which is done regularly in English, though not only with *kinda*. It is likely that this is reason for *kinda* in the following TU as well.



3- FAU (727) ▶ ↻	Occasionally she'd circle back to me clutching some closed-toe prey and say, -This?- and I would try to make an intelligent comment about the shoe, and then finally she bought three pairs and I bought my flip-flops and then as we exited she said, -Anthropologie?-	Ocasionalmente ella regresaba agarrando una víctima de tacón cerrado y decía: -¿Este? -y yo intentaba hacer un comentario inteligente sobre el zapato, y luego finalmente trajo estos tres pares de zapatos, me compró mis sandalias y luego mientras salíamos dijo-: ¿Antropología?
	-I should head home actually,- I said. -I'm kinda tired.-	-De hecho, tengo que volver a casa -dije-, estoy cansada.
	-Sure, of course,- she said.	-Claro, por supuesto -dijo-.

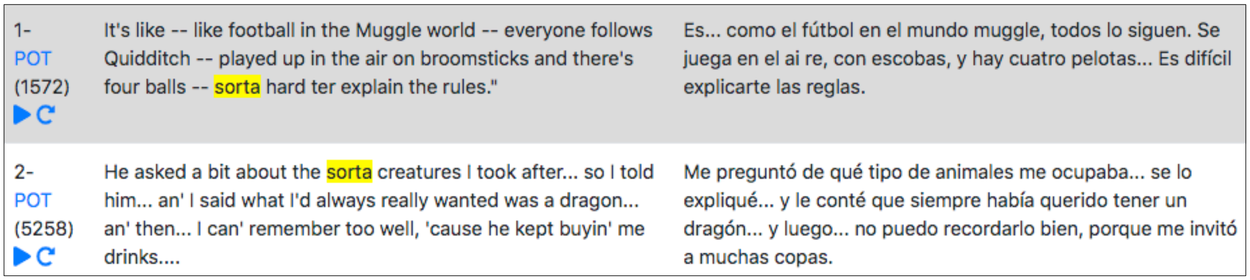
Figure 69. Another example of *kinda* as a way to show politeness.

In this case, the character speaking appears to be attempting to leave a situation he or she may not be too comfortable in but wants to be polite about it. Interestingly, in neither this nor the previous TU is there any reflection of this in the translation.

In another instance, *kinda* is used in the sense of *type of*, as seen in *FAU (1665)*, *…not necessarily this kinda teenagery…*, translated as *tipo de*.

Such literary discourse analysis is beyond the scope of this dissertation, but the cursory glance here is meant to illustrate such possibilities for language learners when working with the data in the corpus.

An examination of *sorta* can be carried out the same way, although the search only yields two results, both from *Harry Potter*.



1- POT (1572) ▶ ↻	It's like -- like football in the Muggle world -- everyone follows Quidditch -- played up in the air on broomsticks and there's four balls -- sorta hard ter explain the rules."	Es... como el fútbol en el mundo muggle, todos lo siguen. Se juega en el ai re, con escobas, y hay cuatro pelotas... Es difícil explicarte las reglas.
2- POT (5258) ▶ ↻	He asked a bit about the sorta creatures I took after... so I told him... an' I said what I'd always really wanted was a dragon... an' then... I can' remember too well, 'cause he kept buyin' me drinks....	Me preguntó de qué tipo de animales me ocupaba... se lo expliqué... y le conté que siempre había querido tener un dragón... y luego... no puedo recordarlo bien, porque me invitó a muchas copas.

Figure 70. Search results for *sorta*.

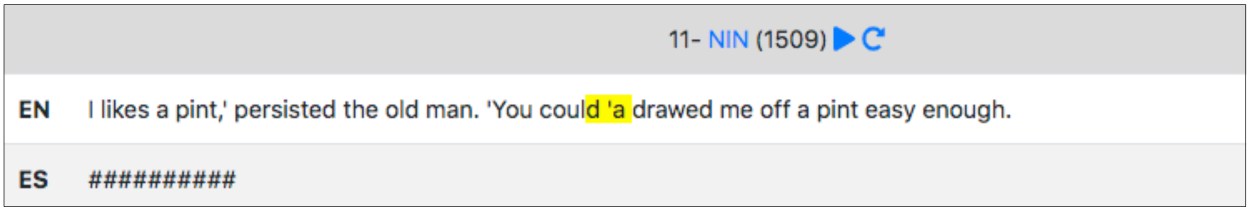
Again, the pronunciation is fairly straightforward, so what learners may find more useful is *how* it is used. In the first TU, there is no equivalent in the translation and it appears to be a form of hedging, in this case to politely say that the rules are too complicated to explain quickly. In the second TU, the equivalent of *sorta* in the translation is *tipo de*, just as how *kinda* or *kind of* may be used to say *type of* as well. These examples can then be compared to the full form of *sort of*. Again, phrases like *kind of* and *sort of* are worth examination by learners as these are common features of spoken English.

Another way in which features of spoken English are represented in dialogue is through the use of apostrophes to represent altered or elided word-final sounds that occur after /n/. A common example of this is when the word-final *g* is replaced with an apostrophe, such as *working* becoming *workin'*, to represent [ŋ] realized instead as [n]. A simple query (displaying 5 per text) can be formulated as ***n'***. This yields 196 results, the majority of which are *an'* in place of *and*, or *-in'* in place of *-ing*. These forms can be heard more clearly before a vowel. In *LOR (505)*, *an' I said…*, the narrator reads it as [ə.nai.sɛd]. In GAT *(2345)*, *One goin' each way…*, the narrator reads it as [wʌn.go.wɪ.nitʃ.wei]. In some instances, *an'* is reduced to a syllabic /n/. This can be heard in *NIN (2650)*, *But the smiles an' the tears…*. This is another case of reduction students need to be familiar with in order to not miss the phonetic cue of the syllabic /n/.

Although fewer in number, there are other results in this search besides *an'* and *-in'*, such as *aren'*, *din'* and if the results were extended to 10 per text, *don'*. In the case of *aren'*, the audio from *LIO (152)*, *Aren' you well?*, is of little help as the narrator palatizes what would normally be /t/ before /j/ in the full form of *aren't you*. As a result, the data is contradictory. Nevertheless, the written form *aren'* still informs students about the possible absence of the word-final /t/ as a spoken cue, as this was an intentional representation on the part of the author for the dialogue. As for *don'* and *din'*, these examples go to show how, much like in *aren'*, the word-final [t] may not be pronounced, and therefore learners should not rely on that cue alone to determine the presence of the negative. In *GAT (866)*, *I din' notice*, the narrator follows the written cues and elides both the second /d/ and final /t/ from what would normally be *didn't*. Furthermore, a blending of /n/ can be heard across word boundaries, realized as [ai.dɪn.no.təs]. The same thing occurs in the case of *don'*. In *POT (918)*, *Yer great puddin' of a son don' need…*, the [t] is elided and [n] is blended across word boundaries.

Similarly, an apostrophe is often used to represent elided word-initial sounds too. This occurs frequently in short, often monosyllabic words. A search query can be formulated two ways, depending on whether we want to include two letters after the apostrophe or one. For only one letter, the query can be formulated as ***\w_ ‘\w_*** and ***\w_ ‘\w\w*** for two letters. It is necessary to have a preceding word to avoid quotations that are set off by single quotes and a space alone is not enough to make sure of that. Additionally, a space must be left at the end so that words longer than one letter do not appear in the search as using ***‘\w**** instead of ***‘\w_*** would still allow quoted phrases to appear rather than reduced forms of words in speech. Also, putting a boundary, such as ***‘\w\b*** would still allow punctuation to appear, such as commas and periods to mark abbreviations, e.g. ***‘Mr.***.

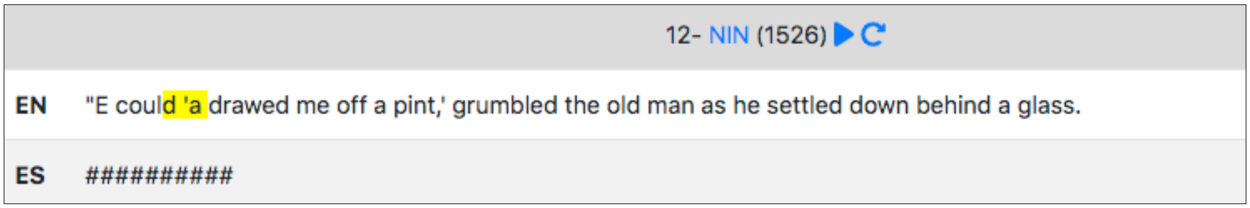
A search for ***\w_ ‘\w_*** yields 16 results, five of which still include apostrophes as quotations. This is impossible to avoid completely as the corpus reads them as the same character. Nevertheless, useful information can be found in the other results. All the TU’s from *Lord of the Flies* show a reduction of *and* to *‘n* whenever naming a pair of people, such as *Sam ‘n Eric* or *me ‘n Simon*. Other results show the same reduction of the auxiliary *have* to schwa as discussed above in the previous section. All instances are from the text *1984* and were not seen previously, as they did not appear in the search results due to the apostrophe. However, these are still examples of phonological cliticization, as heard in *NIN (1509, 1526 and 4176)*:



11- NIN (1509) ▶ C

EN	I likes a pint,' persisted the old man. 'You could 'a drawed me off a pint easy enough.
ES	##########

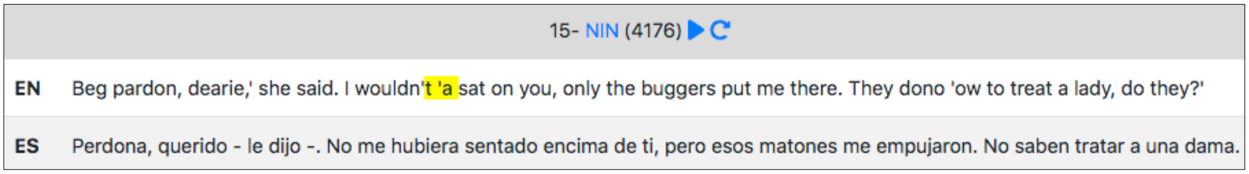
Figure 71a. Examples of *have* reduced to *a* and cliticized.



12- NIN (1526) ▶ C

EN	"E could 'a drawed me off a pint,' grumbled the old man as he settled down behind a glass.
ES	##########

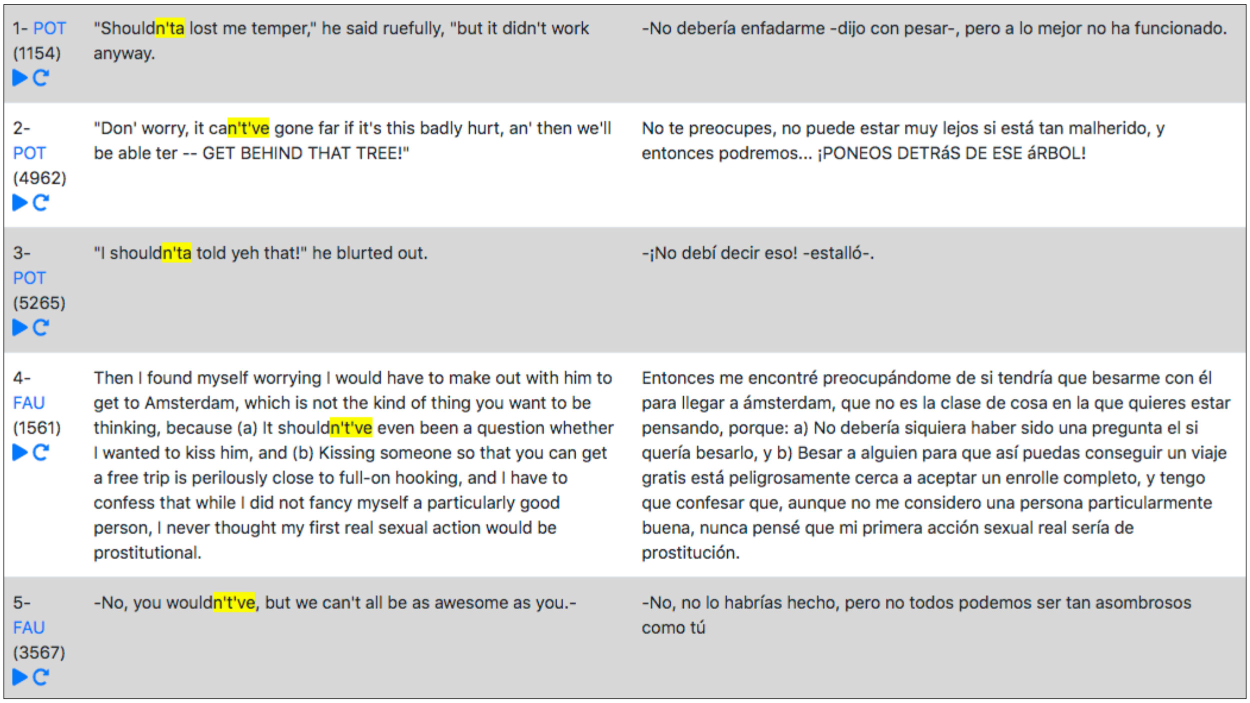
Figure 71b.



15- NIN (4176) ▶ ↻	
EN	Beg pardon, dearie,' she said. I wouldn't 'a sat on you, only the buggers put me there. They dono 'ow to treat a lady, do they?'
ES	Perdona, querido - le dijo -. No me hubiera sentado encima de ti, pero esos matones me empujaron. No saben tratar a una dama.

Figure 71c.

In all three examples, the narrator cliticizes the reduced form of *have*. Of particular interest to students is the third example, in Figure 71c, which involves the negative form. Although not discussed previously, it is possible to add a clitic to a contraction, which itself is a clitic. Other instances of this can be searched in the corpus with a variety of queries. By combining them all, the query ***n't've|n't'a|n'ta*** can be formulated to include all possible versions with the negative contraction. This search yields five results, all from *Harry Potter* and *The Fault in Our Stars*.

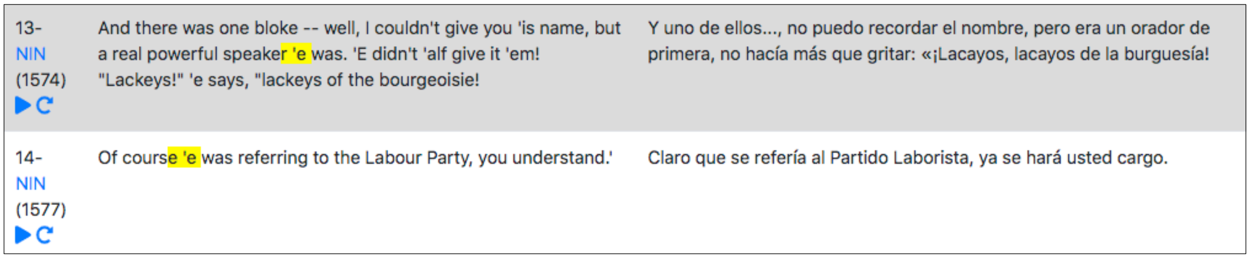


1- POT (1154) ▶ ↻	"Shouldn'ta lost me temper," he said ruefully, "but it didn't work anyway.	-No debería enfadarme -dijo con pesar-, pero a lo mejor no ha funcionado.
2- POT (4962) ▶ ↻	"Don' worry, it can't've gone far if it's this badly hurt, an' then we'll be able ter -- GET BEHIND THAT TREE!"	No te preocupes, no puede estar muy lejos si está tan malherido, y entonces podremos... ¡PONEOS DETRáS DE ESE áRBOL!
3- POT (5265) ▶ ↻	"I shouldn'ta told yeh that!" he blurted out.	-¡No debí decir eso! -estalló-.
4- FAU (1561) ▶ ↻	Then I found myself worrying I would have to make out with him to get to Amsterdam, which is not the kind of thing you want to be thinking, because (a) It shouldn't've even been a question whether I wanted to kiss him, and (b) Kissing someone so that you can get a free trip is perilously close to full-on hooking, and I have to confess that while I did not fancy myself a particularly good person, I never thought my first real sexual action would be prostitutional.	Entonces me encontré preocupándome de si tendría que besarme con él para llegar a ámsterdam, que no es la clase de cosa en la que quieres estar pensando, porque: a) No debería siquiera haber sido una pregunta el si quería besarlo, y b) Besar a alguien para que así puedas conseguir un viaje gratis está peligrosamente cerca a aceptar un enrolle completo, y tengo que confesar que, aunque no me considero una persona particularmente buena, nunca pensé que mi primera acción sexual real sería de prostitución.
5- FAU (3567) ▶ ↻	-No, you wouldn't've, but we can't all be as awesome as you.-	-No, no lo habrías hecho, pero no todos podemos ser tan asombrosos como tú

Figure 72. Results for negative forms with added clitics.

Of the five results, the first three show the reduced form of *have* pronounced as /ə/, despite the second result written as *can't've*, while in the final two, the contracted form of *have* is realized as /əf/.

Returning to the original search for **\w_‘\w_**, one last phenomenon worthy of mention is *‘e* for *he*, written to represent the elided /h/ that is common in pronouns such as *he*, *him*, *his*, *her* and *hers* in spoken English. All three instances in the results are from *1984*. One can be seen at the beginning of the sentence in Figure 71b above, while the other two can be seen in Figure 73 below. The narrator pronounces *‘e* as one would if dictating the letter -*e*-; that is, [i].



13- NIN (1574) ▶ ↻	And there was one bloke -- well, I couldn't give you 'is name, but a real powerful speaker 'e was. 'E didn't 'alf give it 'em! "Lackeys!" 'e says, "lackeys of the bourgeoisie!	Y uno de ellos..., no puedo recordar el nombre, pero era un orador de primera, no hacía más que gritar: «¡Lacayos, lacayos de la burguesía!
14- NIN (1577) ▶ ↻	Of course 'e was referring to the Labour Party, you understand.'	Claro que se refería al Partido Laborista, ya se hará usted cargo.

Figure 73. Reduced forms of *he*.

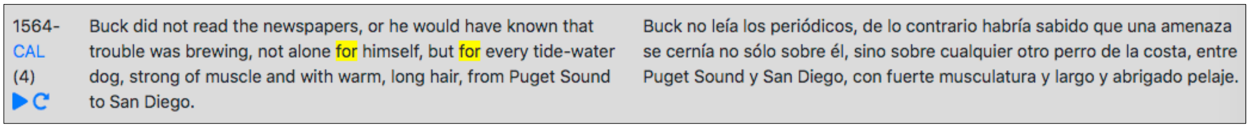
Turning to the other search query, **\w_‘\w\w_** yields 51 results, again with some instances of single quotations instead of an apostrophe. In this case, there is a greater variety of reduced forms that are useful for learners. These are *‘em*, *‘ud*, *‘im*, *‘as*, *‘is* and *‘ad*. *‘Em* can be found in 22 TU’s and is the reduced form of *them*. The data shows it to be pronounced as [əm], with the vowel reduced to schwa. When searched on its own as ***‘em\b***, the results more than double to 46 as it can be found at the end of sentences as well. Unlike previous examples of words realized in their full form at the end of a sentence or tone unit, *‘em* maintains its reduced vowel due to the fact that *them* rarely functions as the focus word (usually the last stressed word, often lexical) of a tone unit.

In the case of *‘ud*, it only appears in four TU’s from a single text, *Lord of the Flies*, and is the reduced form of *would*. It represents speakers’ tendency to reduce its pronunciation to [əd] or [ʊd], eliding the word-initial glide /w/. In *LOR (2498)*, *Then things ‘ud be all right*, the narrator reads *things ‘ud be* as [θɪŋ.zəd.bi:].

In the case of *‘im*, *‘as*, *‘is* and *‘ad*, these four are the result of an elision of the word-initial /h/ in *him*, *has*, *his* and *had*, which is common in spoken English, although in some varieties more than others. All but one example in the previous broad search **\w_‘\w\w_** come from the text *1984* and the narrator elides /h/ in all instances. For this reason, separate individual searches will not yield any additional results, except for *‘im*. A separate search for ***\b’im\b*** yields six results, three more than the original broad search, in which two different pronunciations can be heard, [ɪm] and [əm]. This has the potential to prompt a follow-up activity in which full forms of *him*, *has*, *his* and *had* are explored in the corpus and checked for instances of elision, which there would likely be.

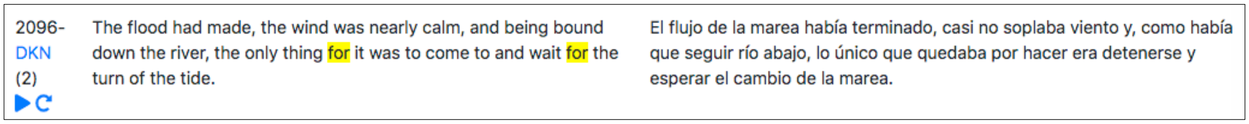
Finally, written representations of how certain prepositions are often realized in dialogue can be found in the data, specifically *of* and *for*, which are both frequently reduced to their weak forms in connected speech. These are represented in the data as *o’* and *fer*. A search for *o’* yields 29 results from three different texts in which it is invariably pronounced as [ə], except in the first result in which *o’* is meant to be *all*, and therefore is pronounced as written, [o]. It is not clear why this is and does not occur elsewhere in the results. A shorter form of *of course*, in this case *o’ course*, appears four times. Then there are certain phrases and expressions that appear regularly in English, such as *cup o’ tea*, *as a matter o’ fact* and *piece o’ cake*. For the most part, however, *o’* frequently follows certain quantifiers such as *some*, *bit*, *one*, *lots* and *couple*. Additionally, some of the cliticized forms discussed previously such as *outta* and *lotta* can be found here as *out o’* and *lot o’*, both with the same cliticized pronunciations as before.

As for *fer*, the query **\b*fer*\b** also yields 29 results, though this time all from a single text, *Harry Potter*. *Fer* is a way of orthographically representing the weak form of *for*, either [fɚ] or [fə]. Both forms can be heard in the corpus. In the third example, *POT (964)*, *…the letter Dumbledore left fer him?*, not only is *fer* realized as [fɚ], but the following word-initial /h/ in *him* is elided, leading to [fə.ɹɪm], even though it is written as *him*. Yet, in *POT (1229)*, *do important stuff fer him*, there is no elision of the word-initial /h/ and so *fer him* is realized as [fɚ.hɪm]. Students can follow this up with a search for **\b*for*\b** (displaying 1 per text) to discover whether or not it is reduced when the narrators read the standard written form. Both forms can be heard throughout the results. In three cases, both the weak and full form can be found within the same TU.



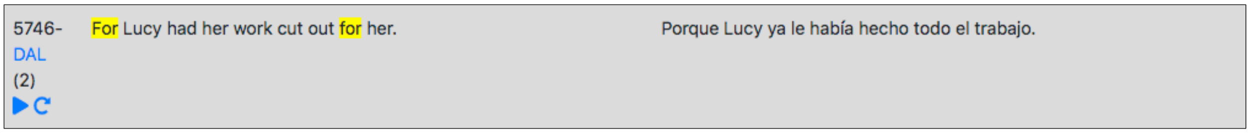
1564- CAL (4) ▶ ↻	Buck did not read the newspapers, or he would have known that trouble was brewing, not alone for himself, but for every tide-water dog, strong of muscle and with warm, long hair, from Puget Sound to San Diego.	Buck no leía los periódicos, de lo contrario habría sabido que una amenaza se cernía no sólo sobre él, sino sobre cualquier otro perro de la costa, entre Puget Sound y San Diego, con fuerte musculatura y largo y abrigado pelaje.

Figure 74a. TU’s in which both the weak and full forms of *for* can be heard.



2096- DKN (2) ▶ ↻	The flood had made, the wind was nearly calm, and being bound down the river, the only thing for it was to come to and wait for the turn of the tide.	El flujo de la marea había terminado, casi no soplaba viento y, como había que seguir río abajo, lo único que quedaba por hacer era detenerse y esperar el cambio de la marea.

Figure 74b.



5746- DAL (2) ▶ ↻	For Lucy had her work cut out for her.	Porque Lucy ya le había hecho todo el trabajo.

Figure 74c.

In the first example from *CAL (4)*, the first instance of *for* is fully realized and the second is reduced, possibly due to the fact that the first is near the end of a tone unit and the second is near the beginning of one. In the case of *DKN (2)*, the first instance of *for* is stressed, possibly for emphasis, while the second appears in the middle of a rapidly spoken tone unit, *…and wait <u>for</u> the turn of the tide*. Finally, in *DAL (2)*, the first instance of *for* at the beginning of the sentence is reduced while the second can be heard in its full form. This may be due to the fact that in the first case, *for* is followed by *Lucy*, a stressed name, and therefore *for* is reduced to avoid stress clashes. As for the second case, it is likely that this is due to the fact that it appears near the end of a tone unit before another function word, as in the first example from *CAL (4)*, which is followed by *himself*. Although beyond the scope of the present work, a detailed analysis can be carried out on the different realizations of *for* in the corpus data. Examining the weak and full forms of function words in the corpus provides students with much needed exposure to such forms, along with the ability to analyze patterns of usage and perhaps even apply them to their own language production.

One way to approach these aspects of conversational English speech in the classroom may be to first ask students how they think they are typically pronounced, and then have them confirm or refute their hypotheses. A more exploratory option is to provide the search queries and asks students what observations they have made. As stated previously, how this data is brought into the classroom depends on many factors and the circumstances surrounding each learning context. Nevertheless, it has been shown here how the representations of English speech in the corpus data can provide useful phonological information for students that may normally get overlooked in already loaded curricula.

3.2.1.8 North American Flap

Consider this: You are a Spaniard who has learned British English for most of your schooling. Now a young adult, you travel to the United States for five weeks to improve your English skills by staying with an American family. After a couple weeks, you are noticing some improvements in your comprehension and picking up on colloquial phrases, such as *sounds good*. One night you are at a noisy bar with one of the members of your American family. You are finally feeling comfortable with the language and decide to order drinks for you and your friend. You order "a beer and a water", but the bartender does not understand the second part of your order. You repeat the word *water* again, but the bartender cannot make out what you are saying even though you are pronouncing as close as you can to how you had learned it so many years ago, /wɔːtə/. After a few more tries, you decide to imitate your American friend's pronunciation and ask for *water*, but now pronouncing the *-t-* as a *-d-*. You feel ridiculous because you never pronounce your English *t*'s as *d*'s as the pronunciation models of your youth rarely presented such a possibility. To your surprise, however, the bartender now comprehends the rest of your order and sets a glass of water next to your beer.

The previous anecdote is a true story of a Spanish friend who had visited the author in his hometown of Milwaukee, Wisconsin. It is meant to illustrate the importance of a common but often overlooked feature of North American English, *the flap*.

Because North American English is so widespread, thanks in large part to the Internet, Hollywood and other media and entertainment industries, it is worth taking a look at such a ubiquitous feature of this variety, which may cause comprehension problems for learners, especially those whose instruction has been based on British models. The phenomenon of the alveolar tap, commonly known as a *flap*, is audible when /t/ and /d/ occur between vowels (Carr, 1999), primarily in North American English, although not strictly limited to this variety (recall the flapping of *gotta* by both British and American English speakers in section 3.2.1.6 above). A flap can occur word-internally or across word boundaries, as long as the VCV sequence is within the same foot. Flapping differs from stops in that there is no sudden burst of air resulting from its realization as there is with stops. Rather, it is a quick movement of the tongue; hence the term *tap*. Furthermore, Whitley (2002) notes that "flapping is most common between a stressed vowel and an unstressed one" and provides the examples of *atom*, [ˋæ.ɾəm], which is often flapped, and *atomic*, [ə.ˋtɑ.mɪk], which is not.

Due to the fact that many of the narrators from the corpus are North American English speakers, it is not difficult to encounter examples of this phenomenon in the data. To locate word-internal flaps, a search for /t/ and /d/ between vowels can be formulated as ***[aeiou][td][aeiou]***, although this will return many word final results such as *made* or *late*. Therefore, it may help to add *\w* to the query, making it ***[aeiou][td][aeiou]\w***. This search still yields many extraneous results that do not provide data on flaps. To work around this, teachers can select specific examples from the initial results that will provide adequate data, such as the word *seated*. Alternatively, a general search can be forgone altogether and teachers can prompt students to search for specific letter combinations likely to yield useful data, such as words that end in V + -*ted*. A separate search for ***\bseated\b*** yields 31 results. When filtered for American English, the number goes down to 12. Eight of the results show flapping, the pronunciation of which is homophonous to *seeded*, [si.ɾəd]. The four results without a flap are from two texts, *Fahrenheit 451* and *1984*, whose professional narrators read with much drama and attention to detail, which may be why [t] is heard instead of a flap. On the other hand, when filtered for British English, all speakers realize the word as [si.təd], showing no flap.

To locate examples of flapping across word boundaries, a general search can be carried out for more concrete instances that can then be explored in separate searches to compare accents. Such a general search (displaying 5 per text) can be formulated as ***[aeiou]t_[aeiou]***. Taking the second result, *PEA (4)*, *…tales that are...*, a separate search for ***\bthat_are\b*** (displaying All results) provides useful data for students. When filtered for American English, all but two of the 27 results show flapping, realizing *that are* as [ðə.ɾɚ] or [ðæ.ɾɚ]. When filtered for British English, all three narrators from the six results fully realized /t/ as [t^h], marked as such for the aspiration that can be heard as /t/ is moved to the onset of the next syllable, *are*, through resyllabification.

Another practical example from the initial search results is when the proceeding word is *it*, as in *HIL (8)*, *…put it on the table*, as *it* itself contains a word-final /t/ which may be flapped as can be heard in this example. This may cause further perceptual issues for learners not accustomed to recognizing the flap as a phonological cue. A broad search involving *it* (displaying 1 per text) may be formulated as ***[aeiou]t_it\b***. When filtered for American English, all but two of the 17 narrators show flapping. In multiple instances, *it* also shows flapping when appearing before another vowel, as in the initial example from *HIL (8)*, as well as in *CAL (70)*, *He could not understand what it all meant,* realizing *what it all* as [wə.ɾɪ.ɾal].

Even longer sequences can be explored by adding more criteria to the search. For instance, the query ***[aeiou]t_it_out_[aeiou]*** yields four results when filtered for American English, one of which, *FAU (1925)*, *I couldn't get it out of…*, flaps each word-final /t/ in *get it out*. Another result, *FAH (448)*, *let it out and…*, flaps *let* and *it*, but the /t/ in *out* is aspirated due to the tone boundary between *out* and *and*. When filtered for British English, only one result is returned, *POT (4938)*, *put it out of its misery*, and does in fact show flapping as well, both in *put* and *it*. However, this time the word-final /t/ in *out* appears to be replaced with a glottal stop rather than a flap. While flapping is possible in certain varieties within Great Britain, it is not as prevalent as in North American English. Nevertheless, this example illustrates the importance of learners at least being aware that such a phenomenon is possible.

Learners are likely to encounter flapping at some point, which means they should have the skills to pick up on such phonological cues, especially when flapping occurs across word boundaries as word-internal flapping is more fixed and unchanging, e.g. *butter* as [bʌ.ɾɚ].

This concludes the present analysis of suprasegmental features in the corpus data. The objective here has been to provide examples of how the suprasegmental phonological data can be exploited for Spanish-speaking learners of English. The co-articulatory phenomena of assimilation, elision, resyllabification, blending, vowel reduction, palatization, cliticization, flapping across word boundaries and the basic notion of the tone unit have been examined in the corpus data from a pedagogical perspective. If these prosodic elements are approached in a way that is appropriate for the given educational context, learners can benefit enormously from the data in the corpus, both in terms of oral comprehension and speech production.

4. FINAL REMARKS

The present work has set out to describe the creation and design of the LITTERA corpus along with potential pedagogical applications within the DDL framework as they pertain to English phonology in ESL contexts. The LITTERA corpus is an audio-textual English-Spanish parallel literary corpus designed to provide authentic examples of audio containing English speech by native speakers. Thanks to the availability of audiobooks, these audio files have been segmented and aligned with the original English language literary texts along with their Spanish translations. The result is a nearly two-million-word corpus that can be accessed freely by anyone with an internet connection.

The corpus is located within two different corpus collections, the CLUVI collection and the SensoGal collection. The version of LITTERA located in the CLUVI collection provides a variety of search capabilities thanks to the availability of regular expressions. The version of LITTERA in the SensoGal collection has been semantically annotated and allows the user to search by lemma, word, or concept, and therefore the ability to explore semantic relationships within Galnet directly from the search results. Nearly all of the pedagogical applications that have been laid forth in the present work were through the CLUVI interface due to the search flexibility provided by the use of regular expressions.

Underpinning the exercises presented here is Data-Driven Learning, an approach to language learning in which students interact directly with the corpus data, thereby discerning patterns and making observations themselves, usually under the guidance of a language teacher. Until recently, very little work has been carried out on how corpora can be used to study phonology by language learners. For this reason, the LITTERA corpus was created and the present investigation undertaken. While the target user is a Spanish-speaking university student, many of the exercises will be relevant to other learners of English.

However, in order to create pedagogically effective material for students, teachers must feel comfortable with using corpora in the classroom. For this reason, video tutorials are available for those who are new to corpora on both of LITTERA's homepages covering topics such as basic corpus features, simple search queries, complex search queries (via regular expressions) and searching by form or by concept. A common criticism from students in the DDL literature is the difficulty in formulating adequate search queries. The tutorials have been designed with such criticism in mind as they are meant to aid new users, both students and teachers alike, in overcoming the initial difficulties surrounding corpus work.

The pedagogical material described in this dissertation revolves around the phonological data in the corpus, particularly the suprasegmental features of English that frequently pose difficulty for Spanish speakers. The goal of the suprasegmental analysis is to point out aspects of connected speech that students should be aware of but are often not explicitly taught and how these phenomena can be examined in the corpus data. No exact recommendations are made because what teachers do with the corpus data will depend on the many factors that make up each individual learning context.

To begin with, the idea of tone units and their boundaries was introduced, as a segment may be altered depending on its location therein. As seen in the subsequent sections, searching for segments at the end of a tone unit tends to produce their "full" or expected forms, while segments not at the end of a tone unit, particularly at word boundaries, are much more likely to be altered by the other neighboring segments. Tone units can be examined in practically any set of search results, as there is at least one tone unit per translation unit.

Following the introduction of tone units, the concept of linking was presented and analyzed in the corpus data through search queries aimed at locating instances of C+V and C+C interactions across word boundaries. The concept of resyllabification was introduced and exemplified through C+V linking. Resyllabification is when a consonant in the coda position becomes the onset of the following syllable. This was examined through phrasal verbs in the data, such as *take off* possibly being misunderstood as *to cough* due to resyllabification of the voiceless stop /k/. Also analyzed was the blending of the same consonant in C+C position across word boundaries. The data showed some degree of blending in all cases except for affricates, which were fully realized in both word-final and word-initial position, although their low frequency in the data indicates that students need not be too concerned with producing such combinations. The concept of blending is useful for students in that it facilitates speech across word boundaries, allowing the speaker to avoid the difficult task of fully pronouncing, say, both a word-final and word-initial /t/.

Upon establishing the basic features of linking, complex consonant clusters across words boundaries were examined in the corpus data. This was done by specifically looking at *st* + C clusters. Multiple DDL questions were proposed throughout this section, beginning with *Is it possible to reduce st + C clusters across word boundaries?*, the answer to which is a clear *yes* judging by the data in the corpus. This question was followed up by others aiming to understand when and why speakers reduce the clusters: *What factors may affect the reduction of the cluster?* and *What leads a speaker to reduce a cluster and what leads a speaker to maintain it?*. The corpus data points to at least three possible causes: the existence of a tone unit boundary, speech tempo and high-frequency expressions such as *la**st n**ight* or *in the fir**st p**lace*. Finally, suggestions are made on how this information and data can be used in the classroom for DDL activities.

Another feature of prosody that can often be found at the convergence of word boundaries is that of assimilation. Assimilation occurs when one segment takes on the characteristics of a neighboring one. As a common aspect of connected English speech, it is no surprise that the corpus contains numerous examples of assimilation across word boundaries. Firstly, instances of /z/ before /ʃ/ in high-frequency verbs such as *is*, *was* and *does* were located in the data, showing how /z/ not only becomes devoiced, but also changes in articulation to that of /ʃ/, making it ambisyllabic. This was followed up with the devoicing of /v/ and /z/ in *have to* and *has to*, in which the phonemes assimilate to the voiceless /t/. The phenomenon of palatization, a type of mutual assimilation, was then discussed and examples were analyzed in the data. It was also observed how certain narrators appeared to palatize purposely for specific characters. The section concluded with a brief look at how the vowel in *the* changes based on the proceeding phoneme being a vowel (except [i]) or a consonant.

After analyzing cases of assimilation, elision and resyllabification, all across word boundaries, the past tense *–ed* morpheme was then examined in the corpus data. This was the only section dedicated explicitly to a single morpheme. While the general rule for its realization—determined by the previous phoneme's voicing or place of articulation—is not particularly complicated, the corpus data shows that the phonetic expression of this morpheme is not as clear-cut as the standard rule would have students believe. This is due to the co-articulatory effects of the aforementioned prosodic phenomenon—linking, blending, resyllabification, assimilation and elision. A very clear set of possible questions for DDL activities aimed at broadening students' understanding of the nuances found in the corpus data are proposed at the end of the section.

The next section departed from the previously discussed prosodic elements and turned to another equally important aspect of spoken English, vowel reduction. Vowel reduction—that is, to a middle vowel, /ə/ or /ɪ/—usually occurs in grammatical words in order to give prominence to lexical words as they are the ones that carry the most meaning. It was shown in the corpus data how modal verbs such as *can*, *have* and *have to* are reduced in fluid speech, although the data also provides instances in which the speaker chooses not to reduce the vowels, either to show contrast or provide emphasis. It is important for students to be aware of this, as the non-reduction of *can*, for example, may lead to confusion with *can't*, which is never reduced. The concept of cliticization was also introduced here, which is when an unstressed word attaches itself to a neighboring stressed word for the sake of rhythm. This was exemplified in the corpus data through the cliticization of *have* and *to*, e.g. *should have* → *shoulda*, *want to* → *wanna*. This phenomenon is made possible thanks to vowel reduction.

After having covered all the main features of English prosody (vowel reduction and the aforementioned array of co-articulatory phenomenon), representations of speech in dialogue were analyzed in the data. Because the corpus is made up of literary texts, there are a plethora of cases in which the author, as well as the narrator in the audiobook, attempts to mimic conversational English. This is one of the biggest advantages of building a speech corpus out of literary texts. Although it is scripted speech, the use of dialogue still provides much insight into different aspects of spoken English, including additional examples of cliticization—particularly the use of contractions, such as *she's gone, there'll*, *it'd*—, word reduction through elision of the initial or final phoneme—*'em*, *'ud*, *'im*, *'as*, *'is*, *'em*, *goin'* and *an'*— and vowel reduction—*for* written as *fer*. Not only can students hear the narrator follow the cues of the author, but they can also see how these features of spoken English are represented in writing.

Lastly, a feature of North American English, the *flap*, also known as a *tap*, was briefly examined in the corpus data. The *flap* occurs when a coronal consonant, particularly /t/ or /d/, appears in an intervocalic position. A *flap* can occur within a word or across word boundaries, the latter causing the most difficulties, as it is less predictable, e.g. ***let it*** *out* and ***put it*** *out*. Furthermore, the *flap* may cause confusion by creating unexpected homophones, particularly for students who are more accustomed to the standard British RP models often used in European curricula, e.g. *seeded* and *seated*. These differences were explored by using the *Spoken English Variety* option when searching the corpus. This allows for exploratory activities to make students aware of *flapping*, especially in high-frequency items such as *that* + V (e.g. *that are*) or *get* + V (e.g. *get up*). While this feature also occurs in other varieties of English, it is easiest to locate examples in North American English, as it is characteristic of this variety.

Throughout many of the sections outlining pedagogical applications, examples were also provided showing how the examination of one phenomenon may lead to more exploratory exercises of other aspects of speech or the text itself. For example, in section 3.2.1.4, what was originally an examination of the data on palatization quickly became a discussion on the possible reasons for the narrator's deliberate use of this feature for only certain characters. This captures the true DDL spirit; becoming intrigued by what is found in the data and hopefully gaining a desire to expand one's use of the corpus beyond the task at hand.

At this point it should be clear that speech corpora, if designed adequately, have the ability to provide language learners with useful data regarding the various features of English prosody, as well as with the tools (e.g. appropriate search options and regular expressions) to locate the examples relevant to their learning needs. Therefore, if such corpora are to find their way into classrooms and other learning contexts, it will likely be due to the fact that they have been designed with that purpose in mind, taking into account the concerns of both the teacher and the student. Through the availability of short tutorial videos and a user-friendly interface, it is the author's hope that users will find much use in LITTERA as a language learning resource.

It is also the author's hope that the present investigation inspires empirical work to be carried out on the study of phonology through corpora in language learning contexts. All the pedagogical applications presented here aim to show what can be done and why it is worth undertaking. While the field of DDL has certainly grown tremendously in recent decades, there clearly still remains much uncharted territory, some of which the work presented here has attempted to illuminate.

List of Tables and Figures

APPENDIX A

Survey questions and results from 28 anonymous participants.
(Top two answers highlighted in bold.)

Question	*Strongly Agree*	*Agree*	*Neutral*	*Disagree*	*Strongly Disagree*
Comprehension (1/10) I am able to watch films and TV shows in English without any subtitles.	**39% (11)**	**46% (13)**	11% (3)	4% (1)	0%
Comprehension (2/10) I am able to follow conversations between native speakers of English.	**39% (11)**	**57% (16)**	4% (1)	0%	0%
Comprehension (3/10) English speech is too fast for me to understand everything.	0%	4% (1)	**25% (7)**	**61% (17)**	11% (3)
Comprehension (4/10) Words seem to disappear in the speech of native speakers.	4% (1)	14% (4)	**36% (10)**	**39% (11)**	7% (2)
Comprehension (5/10) I practice listening to English outside of school.	**64% (18)**	**29% (8)**	4% (1)	4% (1)	0%
Comprehension (6/10) I use YouTube and other streaming websites to practice English.	**79% (22)**	**21% (6)**	0%	0%	0%
Comprehension (7/10) I use music to practice English.	**75% (21)**	**21% (6)**	4% (1)	0%	0%
Comprehension (8/10) British English is easier for me to understand than American English.	4% (1)	11% (3)	**36% (10)**	**29% (8)**	21% (6)
Comprehension (9/10) I can't perceive the differences between different varieties of English.	4% (1)	18% (5)	7% (2)	**46% (13)**	**25% (7)**
Comprehension (10/10) The English classes I've had in primary and secondary school didn't focus enough on oral comprehension.	**50% (14)**	**32% (9)**	11% (3)	4% (1)	4% (1)
Expression (1/6) I become nervous when I have to speak English with a native speaker.	**29% (8)**	**29% (8)**	**29% (8)**	7% (2)	7% (2)
Expression (2/6) I become nervous when I have to speak English with other Spanish speakers.	18% (5)	**21% (6)**	18% (5)	**36% (10)**	7% (2)
Expression (3/6) I am able to express my ideas clearly when I speak English.	18% (5)	**43% (12)**	**29% (8)**	11% (3)	0%
Expression (4/6) The English classes I've had in primary and secondary school didn't focus enough on oral expression.	**46% (13)**	**46% (13)**	4% (1)	4% (1)	0%

Expression (5/6) I try to speak English with a specific accent (e.g. North American, British, Australian, etc.).	18% (5)	**36% (10)**	**21% (6)**	18% (5)	7% (2)
Expression (6/6) I worry about making mistakes when I speak.	**36% (10)**	**54% (15)**	4% (1)	7% (2)	0%
Language Experience (1/17) I have spent more than 3 months at a time in an English speaking country.	7% (2)	11% (3)	0%	**14% (4)**	**68% (19)**
Language Experience (2/17) I have regular contact with native speakers of English.	11% (3)	7% (2)	21% (6)	**36% (10)**	**25% (7)**
Language Experience (3/17) I frequently practice English with native speakers.	14% (4)	4% (1)	21% (6)	**32% (9)**	**29% (8)**
Language Experience (4/17) I have had private English teachers.	**32% (9)**	18% (5)	7% (2)	4% (1)	**39% (11)**
Language Experience (5/17) I have attended a language academy.	**43% (12)**	14% (4)	0%	4% (1)	**39% (11)**
Language Experience (6/17) I grew up watching films dubbed in Spanish.	**46% (13)**	**36% (10)**	11% (3)	4% (1)	4% (1)
Language Experience (7/17) I prefer to watch an English-language film dubbed in Spanish.	0%	0%	**18% (5)**	14% (4)	**68% (19)**
Language Experience (8/17) I prefer to watch an English-language film in English but with Spanish subtitles.	**21% (6)**	**21% (6)**	18% (5)	11% (3)	**29% (8)**
Language Experience (9/17) I prefer to watch an English-language film in English and with English subtitles.	**43% (12)**	**36% (10)**	14% (4)	0%	7% (2)
Language Experience (10/17) I prefer to watch an English-language film in English and without subtitles.	18% (5)	**39% (11)**	**25% (7)**	18% (5)	0%
Language Experience (11/17) I would read more literature in English if I could easily consult the translation.	7% (2)	**25% (7)**	**29% (8)**	**25% (7)**	14% (4)
Language Experience (12/17) When I read in English, I look up every word I don't know.	**18% (5)**	**39% (11)**	21% (6)	**18% (5)**	4% (1)
Language Experience (13/17) Knowing another language besides Spanish or Galician has helped me with English.	7% (2)	**25% (7)**	**39% (11)**	21% (6)	7% (2)
Language Experience (14/17) Much of my English learning is done online or through digital media.	**50% (14)**	**21% (6)**	14% (4)	14% (4)	0%
Language Experience (15/17) I use social media to practice English.	**50% (14)**	**36% (10)**	11% (3)	4% (1)	0%
Language Experience (16/17) I know what a corpus is (in Linguistics).	**18% (5)**	14% (4)	14% (4)	**36% (10)**	**18% (5)**
Language Experience (17/17) I have used a corpus for learning English.	4% (1)	11% (3)	**29% (8)**	**39% (11)**	18% (5)

APPENDIX B

Transcript from the video tutorial series for LITTERA in CLUVI

Part 1 – Corpus Overview

This video will present an overview of some of the basic features and characteristics of the LITTERA corpus. The LITTERA corpus is an audio-textual English-Spanish parallel literary corpus. In other words, it contains two types of data format, text and audio. It is a bilingual corpus — the texts and their translations have been aligned at the sentence level. Its domain is English language literary works. By using literary texts, we have been able to add the audio from the corresponding audiobooks to the segmented and aligned text.

The corpus is actually located as a sub-corpus in two different corpora, the CLUVI corpus and the SensoGal corpus, both hosted at the University of Vigo in Vigo, Spain. The CLUVI corpus is a collection of parallel corpora with Galician as the central language, although it contains a variety of other language pairs as well. The SensoGal corpus is a collection of semantically annotated parallel corpora, many of which are taken directly from the CLUVI corpus. Although the exercises in this video are specifically intended for working with the LITTERA sub-corpus, many of the basic skills gained here can also be applied to the other sub-corpora within CLUVI. Instructional videos for searching LITTERA via the SensoGal corpus are available on LITTERA's corresponding homepage in SensoGal.

This three-part tutorial will consist of a series of short exercises aimed at drawing the user's attention to some of the different features and search capabilities of the corpus. Each exercise question will be shown on the screen, at which point the user should pause the video and attempt to answer the question. Then un-pause the video to see the solution.

Part 1 will provide a general overview of some of LITTERA's characteristics. Part 2 will focus on simple search queries, and part 3 will focus on complex search queries as LITTERA provides a range of search options thanks to the use of regular expressions.

So let's begin with the exercises. If you are accessing this video from the LITTERA homepage, you are ready to begin. You may choose your language preferences in the upper right hand corner. This video will be using the English version of the corpus. If you are watching this video from a different source, you can find the LITTERA home page at the link on the screen.

1. How many total words are in the corpus? How many per language?

Explanation: There are almost 2 million total words in the corpus, 1,968,676 to be exact, 983,618 in English and 985,058 words in Spanish. It is important to remember that corpora come in all different shapes and sizes. Some have millions of words while other local DIY corpora may only have a few thousand. Corpus size will depend on many factors, such as research objectives, text availability, legal licensing, cost, human resources, and so on. In the case of LITTERA, many of the texts included in the corpus have been taken from the English philology program at the University of Santiago de Compostela, except, of course, children's literature and young adult fiction, which were included to provide a wider range of registers.

2. According to the homepage, how many translation units are in the corpus?

Explanation: The answer is 63,508. As we will see later on, a translation unit is a bilingual pairing that contains the same information in different languages.

3. What is the three-letter reference code (found in brackets) for Virginia Woolf's *Mark on the Wall?* And for Ernest Hemingway's *The Old Man and the Sea?*

Explanation: The reference code for Virginia Woolf's *Mark on the Wall* is *MRK*, while the code for *The Old Man and the Sea* is *OLD*. These will be important later on when we search the corpus.

4. How many English words are in *Lord of the Flies*? How many translation units?

Explanation: There are 59,128 words in *Lord of the Flies* and 5,555 translation units.

5. Which book contains the most words in English? The least?

Explanation: The book with the most words in English is *Sense & Sensibility* by Jane Austen — 118,503. The book with the least is *The Very Hungry Caterpillar* by Eric Carle — 225 words. This exercise is meant to bring up the idea of representativeness in a corpus. Even though children's books are included in the corpus, because of their brevity and small number — there are only three — they make up less than 1% of the total number of words in the corpus.

6. Which book contains the most translation units?

Explanation: The book with the most translation units is *The Hunger Games* by Suzanne Collins — 6,628. Note that the book with most words, *Sense & Sensibility*, does not have the most translation units. In fact, it has less than *Lord of the Flies*, which contains almost half as many words as Sense & Sensibility, which brings us to our final question.

7. Why might the book with the most words not have the most translation units?

Explanation: This is because, on average, *Sense & Sensibility* contains many more words per translation unit than, say, *Lord of the Flies*. We can see this difference by looking at each individual text via the Audiobooks option. The difference in terms of numbers of words per translation unit is quite clear.

Now that you have a general understanding of the corpus, go to the next video for part 2 to begin the search exercises.

Part 2 – Simple Search Queries

Welcome back. Now that we have a general idea about the structure of the corpus, it's time to start searching and discovering the range of features that are available.

Turning our attention to the search menu, let's first have a look at the different options available. We can search in English, in Spanish, or both simultaneously. The *Lexical Equivalents* option, which is automatically checked by default, will allow us to search by lemma and will highlight the corresponding equivalent in the translation. The Wider Context option allows us to see the context surrounding each of the results. We can control the number of results with the *Results* drop down menu. Lastly, we can select the spoken variety of English we want, British or American. Let's begin the search exercises.

1. Without adjusting any of the default settings, how many total results are there for *play*?

Explanation: There are 155 total results, although because the default is 20 and we did not adjust it, we can only view 20 translation units. This is a good moment to draw your attention to a few other features of the corpus. At the top, we can find what appears to be a dictionary entry for *play*, with a list of translation equivalents in Spanish. These equivalents are based on the word sense entries for Galnet, a semantic dictionary for Galician based on WordNet, which also includes a variety of other languages. The entries in Galnet only include lexical words; that is, nouns, verbs, adjectives and adverbs. Therefore, a preposition or an article will not yield any lexical equivalents. For more on Galnet, see the SensoGal homepage where you can find tutorials similar to this one. Returning to the search results page, we can see the lexical equivalents highlighted in green. Now locate the three-letter code accompanying each translation unit. By clicking on this link, we can refer back to the corpus bibliography, and we can return to the results page via the link at the top. Next to the three-letter code, is the index number of that translation unit within the book it comes from. The first translation unit, from *The Pearl*, is the 1,262nd translation unit in that book. Next to that index number are the audio options. By clicking play, we can listen to the audio. The rewind button goes back two seconds so that a segment of the audio can be repeated as many times as needed. Lastly, we can change the display format of the translation units from the default horizontal to vertical by clicking the icon in the upper right. Now let's return to the search menu by clicking *search again*.

2. Uncheck the *Lexical Equivalents* option. How many total results are there for *play* now? Why is this number different than the previous search? (Feel free to adjust the Wider Context option or the number of results displayed. However, do not adjust the Spoken English Variety option as will omit all results in which the spoken variety is not the selected one.)

Explanation: The total results are now 423. Why is this? Well, as we can see, without the *Lexical Equivalents* option checked, the corpus returns any word containing the string of characters *p-l-a-y* instead of treating *play* as a lexical entity. This is why we can see words like *played*, *display* or *playing*, as well as *play*. Also note, the Galnet dictionary entries are no longer present at the top of the page, nor are any words highlighted in green.

3. Leaving the English search box empty, but checking the *Lexical Equivalents* option, how many results are there if we search for *obra* in Spanish?

Explanation: The answer is 53. Note the dictionary entry at the top of the page again. Note also, that only some of the words are highlighted in green. This occurs when none of the words in the dictionary entry shown at the top of the page appear in the English text. Therefore, the corpus finds no direct equivalent. Also note translation unit number four. *Word* is highlighted in green when it should be *deed* that is highlighted. A similar mistake is made in translation unit number five with the word *worn*. Because these words don't appear in the dictionary entry at the top, the corpus must have mistaken these words for *work*, which *is* an entry above. Finally, note that no other variants of *obra* turned up in the search. Only *obra*.

4. Repeat the same search as in exercise 3, but now without the *Lexical Equivalents* option checked. How many results are there?

Explanation: Now there are 187 results, including words like *sobra*, *obras*, *cobrar* and *maniobras*.

5. Now search for *obra* in Spanish and *play* in English simultaneously. How many results are there? (Note: when both languages are searched simultaneously, the *Lexical Equivalents* option no longer applies, whether it is checked or not.)

Explanation: There are 12 results. Note how the terms in both languages are highlighted in yellow. Again, no dictionary appears at the top of the page because the *Lexical Equivalents* option can only be applied if the search is carried out in one language.

6. Follow the menu options so that you are only searching in George Orwell's 1984 (Full-text search → Audiobooks → English-Spanish → George Orwell – 1984). Leaving the *Lexical Equivalents* option checked, how many times does the word *party* appear in *1984*? How many times does it appear in the entire corpus?

Explanation: Party appears 259 times in *1984* and 425 times in the corpus as a whole. The option to search a book individually allows the user to get a better understanding of how language is used by an author in a specific text, among other things.

This concludes part 2. In this video we looked at some of the basic search options available to explore the corpus. In part 3 we will see how more complex searches can be carried out using regular expressions.

Part 3 – Complex Search Queries Using Regular Expressions

In this video, we are going to introduce some basic regular expressions, which can make a search within LITTERA more flexible. Regular expressions are characters that allow the user to search for multiple items in a single query. The CLUVI corpus offers a variety of regular expressions to search LITTERA.

All the regular expressions are explained on the *Help* page. However, in this video, we are going to look at a small handful of some of the most common regular expressions through a series of practical exercises.

1. Place \b before the word *play* (without a space in between \b and the word) and carry out a search. Judging by the results, what does \b do?

Explanation: A backward slash and *b* create a word boundary. Therefore, words like *display* do not appear in the results because we have placed a boundary there. However, words like *played* or *playing* will still appear in the results because we haven't set a boundary at the end of the word. We could also put a space before the word instead of a boundary, but that won't count any words at the beginning of the translation unit or a quote. Therefore, a boundary is placed to include all instances, regardless of what precedes it.

2. What do brackets [ ] do judging by the results from the query *ma[kd]e sure*?

Explanation: In this case, brackets return all instances of *make sure* and *made sure*. The brackets search for zero or one instance of each character within them. We could add an *r* to the brackets, but that would not change the results as there are no instances of *mare sure* in the corpus, so the program ignores that and returns all instances of *make sure* and *made sure*. This is a useful regular expression for irregular verbs in the past tense, such as *make*.

3. What does a dash – do judging by the results from the query *[b-d]ed\b*?

Explanation: The dash gives a range, in this case from b to d; that is, b, c and d.

4. How can we elaborate the previous search, *[b-d]ed\b*, with dashes to return all instances where a consonant appears before the past tense *ed* morpheme (excluding w and y)?

Explanation: We can form the query as it appears on the screen: *[b-df-hj-np-tvxz]ed\b*. With this query we simply skip the vowels, along with y and w. This is a useful search to study the pronunciation of the past tense morpheme as it is the preceding phoneme that determines its phonetic realization.

5. What does a vertical pipe | do judging by the results from the query *make sure*|*made sure*?

Explanation: The vertical pipe translates to *or*. Therefore, the query reads as *search for all instances of make sure* **or** *made sure*. This returns the same number of results as the query with brackets from exercise 2 as it is just another way to carry out the same search. Much like in computer programming where there are often multiple ways to write code to get the same end result, there are often a number of ways to search the corpus using regular expressions that will return the same or very similar results.

6. What does the question mark ? do judging by the results from the query *\bci?eg* in the Spanish search box?

Explanation: The question mark returns zero or one instance of the preceding character. This is why *ciego* and *cegarlo* both appear in the search results.

7. How can I return all possible forms of stem-changing verb *negar* (*negando*, *niego*, *negó*, *negué*, *negaron*, *niegue*, etc.) while minimizing any unwanted results? (Note: this may take a few attempts to find the most optimal search.)

Explanation: To return all possible forms, we can formulate the query as seen on the screen: *\bni?eg[aóu]*|*\bniego\b*. Had we placed the unaccented *o* from *niego* within the brackets, various forms of *negociar* would overwhelm the results. Because *niego* is the only instance requiring an unaccented *o*, we can separate it with an *or* statement to eliminate all the extra results pertaining to *negociar*. However, despite refining the search in such an efficient way, we can see that unexpected words still show up in the results as well, such as *negativa* and *negativas*. These are not as easily filtered out. But because they are minimal, they do not hinder our corpus search like the forms of *negociar* would have.

8. How many results are there if we want to return all forms of *negar* when they correspond to the English verb *shake* (past form: *shook*)?

Explanation: There are 30 results. We can use the same formula from the previous exercise as seen on the screen (*\bni?eg[aóu]|\bniego\b*), but now we must also search for all forms of shake. To do so, we can use the vertical pipe and formulate the query as *shook|shak*. It is not necessary to add the *e* because this way we are also including the gerund *shaking*.

9. What does the backward slash \ do judging by the results from the query *\?* or *\[* ?

Explanation: The backward slash takes a character that normally functions as a regular expression and makes it literal. So here we are searching for all instances of question marks and brackets. When searching \?, remember that the number of translation units does not equate to the number of questions in the corpus as there may be more than one question in a translation unit. Finally, notice how with certain letters this does the opposite, such as *\b* to mark a boundary. With *\b*, we are taking a letter and applying it as a regular expression.

10. What does *\w* do judging by the results from the queries *\b\wat\b* and *\b\w\wat\b*?

Explanation: This regular expression returns any character that can form part of a word. That is why the results in the first query include words like *eat*, *mat*, *pat* and *fat*, and in the second query we can find words like *that*, *goat* and *beat*. The only thing these words have in common is the number of letters designated by the search query. Therefore, when using *\w*, character options such as punctuation marks are automatically discarded.

11. What does adding an asterisk after *\w* do in the query *\b\w*at\b*?

Explanation: An asterisk means one or more of the previous character, in this case *\w*. Here, the search returns any word that ends in *a-t*. However, if we're looking for a word that ends in *a-t*, another way to search for this would be to simply put nothing before *a* and put a boundary after *t* as you can see on the screen.

12. How can we search for all sequences of three words in which the third word is always *up* (e.g. *give it up* or *he looked up*)?

Explanation: The simplest way to indicate separate words is to use \w* separated by spaces. Therefore, the search could look like the following *\w* \w* up\b*. This is very useful when looking for sequences involving multiple words, such as English phrasal verbs. As we can see from the results, searching by particle is a way to locate phrasal verbs in the corpus.

This concludes part 3 and the series of video tutorials for the LITTERA corpus within CLUVI. For further information on the regular expressions that are available, we encourage you to go to the *Help* page, the link to which is located in the header. There you can find all the different regular expressions with examples of how to use each one. We hope you find LITTERA to be a useful resource for whatever your research objectives may be.

APPENDIX C

Transcript from the video tutorial series for LITTERA in SensoGal

Part 1 – Corpus Overview

This video will present an overview of some of the basic features and characteristics of the LITTERA corpus in SensoGal. The LITTERA corpus is an audio-textual English-Spanish parallel literary corpus. In other words, it contains two types of data format, text and audio; it's bilingual; the texts and their translations have been aligned at the sentence level; and its domain is English language literary works; by using literary texts, we have been able to add the corresponding audio from audiobooks to the segmented and aligned text. The corpus is actually located as a sub-corpus in two different corpus collections, the CLUVI corpus and the SensoGal corpus, both hosted at the University of Vigo in Vigo, Spain. The CLUVI corpus is a collection of parallel corpora with Galician as the central language, although it contains a variety of other language pairs as well. The SensoGal corpus is a collection of semantically annotated parallel corpora, many of which are taken directly from the CLUVI corpus. The exercises in this video are specifically intended for working with LITTERA via the SensoGal interface. Instructional videos for LITTERA via CLUVI are available on LITTERA's corresponding homepage in the CLUVI corpus.

This three-part tutorial will consist of a series of short exercises aimed at drawing the user's attention to some of the different features and search capabilities of the corpus in SensoGal. If you would like to use a different corpus in SensoGal, this tutorial will still be of use as it covers all the basic search features even when accessing SensoGal through other corpora. Each exercise question will be shown on the screen, at which point the user should pause the video and attempt to answer the question. Then un-pause the video to see the solution.

Part 1 will provide a general overview of some of LITTERA's characteristics. Part 2 will focus on searching by form, and part 3 will focus on searching by concept.

So let's begin with the exercises. If you are accessing this video from the LITTERA homepage in SensoGal, you are ready to begin. You may choose your language preferences in the upper right hand corner. This video will be using the English version of the corpus. If you are watching this video from a different source, you can find the LITTERA home page at the link on the screen.

1. How many total words are in the corpus? How many per language?

Explanation: There are almost 2 million total words in the corpus, 1,968,676 to be exact, 983,618 in English and 985,058 in Spanish. It's important to remember that corpora come in all different shapes and sizes. Some have millions of words while other local DIY corpora may only have a few thousand. Corpus size will depend on many factors, such as research objectives, text availability, legal licensing, cost, human resources, and so on. In the case of LITTERA, many of the texts included in the corpus have been taken from the English philology program at the University of Santiago de Compostela, except for children's literature and young adult fiction, which were included to provide a wider range of registers.

2. According to the homepage, how many translation units are in the corpus?

Explanation: The answer is 63,508. As we will see later on, a translation unit is a bilingual pairing that contains the same information in different languages.

3. What is the three-letter reference code (found in brackets) for Virginia Woolf's *Mark on the Wall?* And for Ernest Hemingway's *The Old Man and the Sea?*

Explanation: The reference code for Virginia Woolf's *Mark on the Wall* is *MRK*, while the code for *The Old Man and the Sea* is *OLD*. These reference codes will be important later on when we search the corpus.

4. How many English words are in *The Lord of the Flies*? How many translation units?

Explanation: There are 59,128 words in *Lord of the Flies* and 5,555 translation units.

5. Which book contains the most words in English? The least?

Explanation: The book with the most words in English is *Sense & Sensibility* by Jane Austen — 118,503. The book with the least is *The Very Hungry Caterpillar* by Eric Carle — 225 words. This exercise is meant to bring up the idea of representativeness in a corpus. Even though children's books are included in the corpus, because of their brevity and small number — there are only three — they make up less than 1% of the total number of words in the corpus.

6. Which book contains the most translation units?

Explanation: The book with the most translation units is *The Hunger Games* by Suzanne Collins — 6,628. Note that the book with most words, *Sense & Sensibility*, does not have the most translation units. In fact, it has less than *The Lord of the Flies*, which contains almost half as many words as Sense & Sensibility. This brings us to our final question.

7. Why might the book with the most words not have the most translation units?

Explanation: This is because, on average, *Sense & Sensibility* contains many more words per translation unit than, say, *Lord of the Flies*. We can see this difference by looking at each individual text via the Audiobooks option. The difference in terms of numbers of words per translation unit is quite clear.

Now that you have a general understanding of the corpus, go to the next video for part 2 to begin the search exercises.

Part 2 – Search by form

Welcome back. Now that we are familiar with the composition of LITTERA, it is time to turn our attention to the search menu on the left. Because all the corpora in SensoGal are both lemmatized and semantically annotated, we have two different ways to search the corpus; that is, by form or by concept. This is evident from the search menu. In this part, through a series of practical exercises, we will explore the available features of LITTERA in SensoGal when searching by form.

As you can see, searches can be carried out in English, Spanish or both simultaneously. The *Search Lemma* box is checked by default. If desired, the user can specify the search term's part of speech. However, because the corpus is tagged semantically, this is limited to lexical words, i.e. nouns, verbs, adjectives and adverbs. Finally, the user can choose a few more basic options, such as whether or not to show the surrounding context in the results, setting the number of results, or specifying the variety of spoken English. Now that we are familiar with the search options, let's begin the exercises.

1. Uncheck the *Search Lemma* box but leave all the other default settings as they are. Type the word *show* into the English search box and carry out a search. How many total translation units are found (not necessarily shown) and how many equivalents were not aligned at semantic level (see information located above the first result near the top of the page)?

Explanation: This search yields 178 translation units and 15 equivalents not aligned at the semantic level. A translation unit refers to aligned bilingual pairs, in this case in English and Spanish. These are numbered in the results. Within each translation unit, the user can see the original text, in black, and the semantically annotated text in blue. The queried word, then, appears highlighted in yellow and is accompanied by an ILI code. ILI stands for Interlinguistic Index. By clicking on this code, we are redirected to the Galnet page that corresponds to a specific concept. Galnet is a multilingual "sense" dictionary for Galician based on WordNet. The concept for the example we clicked on is "make visible or noticeable" and accompanied by other words in other languages that correspond to that same concept. Therefore, when equivalents are aligned at the semantic level, words in the two languages share the same sense or concept according to Galnet, as can be seen in the first translation unit. *Show* and *mostraran* correspond to the same sense or concept — they share the same ILI code. However, in the second translation unit, there is no lexical equivalent in Galnet found in the Spanish translation. If we only want to view those translation units in which equivalents are found in the other language, we can click on the funnel icon in the upper right, like so.

Along with the results number, each translation unit contains additional information. Locate the three-letter code accompanying each translation unit. By clicking on this link, we can refer back to the corpus bibliography, and we can return to the results page via the link at the top. Next to the three-letter code, is the index number of that specific translation unit within the book it comes from. The first translation unit, from *The Pearl*, is the 230th translation unit from that text. Next to *that* index number is a set of arrows. By clicking on the downward arrow, all the available ILI codes are visible in the translation unit. The upward arrow hides them again. Next to the arrows are the audio options. By clicking play, we can listen to the audio. The rewind button goes back two seconds so that a segment of the audio can be repeated as many times as needed. Lastly, we can change the display format of the translation units from the default horizontal to vertical by clicking the icon in the upper right. Now let's return to the search menu by clicking *search again*.

2. Check the *Search Lemma* option and carry out a new search for *show*. How many total translation units are there?

Explanation: There are 401 examples when searching for *show* by lemma. Note that in the previous exercise when searching by *word*, only a single instance of *show* was returned. When searching by lemma, all forms of *show* are returned; that is, different tenses, gerunds and participles. Note the dictionary entries listed at the top of the page. These are all examples that coincide in the same concepts as *show* in GalNet. Furthermore, it is now the lemma that is highlighted rather than the actual word in the text as in the previous search by word.

3. Clear the English search box and type *mostrar* into the Spanish search box (maintain the *Search Lemma* option). How many total search results are there?

Explanation: There are 258 total results. Note the GalNet dictionary entry at the top of the page, this time with English equivalents instead of Spanish.

4. How many total results are there when searching for *mostrar* and *show* simultaneously (keeping *Search Lemma* checked)? After filtering out those results without lexical equivalents (via the funnel icon), what is the GalNet concept for the second result (PEA – 715, numbered as 4)?

Explanation: There are 104 total results. Notice that the search returned the lemmas for both Spanish and English. We can see infinitives and conjugated forms of each verb. By clicking the ILI code of the second matched result, we are taken to GalNet where we can find the concept described as "give expression to". If we scroll down, we can get a better idea of how the senses, or concepts, in GalNet are organized; that is, through a hierarchical structure of *hypernyms* and *hyponyms*. In this instance for *show*, the corresponding hypernym is *make known; pass on, of information*. *Hypernyms* are more general and encompassing than hyponyms. For example, a hypernym of *blue* might be *color* since color is a more general term. Therefore, *blue* is a hyponym of *color*. We can see that the hyponyms here are more specific than the original concept, with such examples as *express through a scornful smile* or *express or state indirectly*. These relationships can be visualized more clearly by clicking on the *show image* option. Because GalNet is based on WordNet, a semantic dictionary originally developed in the United States at Princeton University, all of the sense descriptions for *hypernyms* and *hyponyms* are in English.

5. Repeat the simultaneous search for *mostrar* and *show* but uncheck the *Search Lemma* option. How many total results are there now?

Explanation: This search returned only 21 total results. Because we searched by word and not by lemma, the query returned only a single form of each word; that is, *show* and *mostrar,* and no other forms or conjugations. Note that now the actual words from the text are highlighted in yellow instead of the corresponding lemmas beside them.

6. Set the number of results to All. Now search for *show* by lemma (leaving the Spanish box empty), but specifying *noun* as the part of speech. What is the ILI code for the only two semantically aligned equivalents?

Explanation: The information at the top tells us that there are only two equivalents that were successfully aligned. We can find them by using the funnel icon. The ILI code can be seen here on the screen. In both cases, *show* corresponds to *espectáculo* in the translation. However, if we look back at the results without equivalents, we can see cases where *espectáculo* is in the translation, yet it is annotated with a different concept. Such is the case for the third result. When working with corpora that have been annotated automatically, one must keep a critical eye on the results to avoid reaching inaccurate or incomplete conclusions.

7. Search for *adult* by lemma in English. In which variety of spoken English is the first syllable stressed? In which variety is the second syllable stressed?

Explanation: In British English, the first syllable is stressed, as in *adult* while in American English it's the second syllable, *adult*.

This concludes the part 2 of this tutorial series for the LITTERA corpus in SensoGal. In this section we have seen how to search by word or by lemma, and how to cautiously interpret the annotated information. Semantic tagging allows us to understand the underlying concept, rather than just general translations. Because the same word can take on different meanings, and therefore different senses or concepts depending on the context in which the word is used, this type of information is of much relevance to the user.

Part 3 – Search by concept

In part 3 of this tutorial series, we will see how users can search by concept in SensoGal using the LITTERA corpus. In part 2, we saw how the corpora in SensoGal are lemmatized and semantically annotated and that each lexically annotated word corresponds to a concept in Galnet, the Galician WordNet. Instead of searching by word, in this video we will search by concept to gain a better understanding of SensoGal and Galnet.

In order to search by concept, we need to use the ILI code. As established in part 2, ILI stands for Interlinguistic Index. To do so, we can either go directly to Galnet, or we can locate a given word's ILI code from search results when searching by form. In the following exercise, we are going to do the latter.

1. Search the word *act* by lemma. What is the ILI code for *act* in the first translation unit?

Explanation: The ILI code for *act* in the first translation unit can be seen on the screen (ili-30-02367363-v). Note that all ILI codes begin the same way, ILI-30, followed by a series of numbers which identify the concept, along with the part of speech, in this case a verb, hence the *v*.

2. Click on the ILI code for *act* in the first translation unit. What is the glossary definition for this concept (in English)? What are the Spanish equivalents for this concept?

Explanation: The concept is *to perform an action, or work out or perform (an action)*. The Spanish equivalents are *actuar*, *hacer*, *llevar a cabo* and *obrar*.

3. Returning to the search results, copy the ILI code for *act* from the first translation unit and paste it in the English box in the *Search by concept* section of the search menu. How many total results are there?

Explanation: There are 74 total results from this search. It would appear that all the results in English correspond to *act*, despite *move* also being an option according to the GalNet entry for this concept.

4. Now search for the same concept, but only in Spanish. How many total results are there? Why are the results different?

Explanation: Now there are 208 search results. The results here are different because Spanish has more words that correspond to this concept than English does in GalNet. Furthermore, one must take into account that this corpus, like all corpora, has a specific size and therefore certain limits, which is part of why it failed to locate instances of *move* for this concept. But what if we repeat this search, only this time adding a different concept in English than the one in Spanish?

5. The first result from the previous search did not contain semantic equivalents. Search for the concept corresponding to *work* in the first English translation unit (ili-30-02413480-v) while simultaneously searching for the same concept from the previous search (ili-30-02367363-v) in Spanish. How many different books yield results for this search?

Explanation: This search yields results from three different books: *The Pearl*, *The Call of the Wild* and *1984*. We can compare the two concepts by the terms and their glossary definitions. If we click on the concept corresponding to *work* in the search results, we can see that the glossary definition is to *exert oneself by doing mental or physical work for a purpose or out of necessity*. The words in Spanish that appear are *esforzarse* and *trabajar*, while in English there is only the word *work*. We can compare this and the other words to the previous concept for *act*. The co-appearance of these two concepts in a translation unit then is not surprising since *carrying out an action* is also *doing work*. The fact that these appear similarly in different books is not an accident and we can infer from this that these two concepts are semantically similar.

Like the majority of Natural Language Processing, or NLP technologies, semantic annotation is far from being an exact science and in many ways incomplete, even more so for languages other than English. Nevertheless, it is clear that the search tools available in SensoGal are of great value, both in language pedagogy and linguistic research. This concludes part 3 of our tutorial for exploring LITTERA through SensoGal. We hope you find this to be a useful resource in whatever your language interests may be.

REFERENCES

Amador-Moreno, C. P. (2010). How can corpora be used to explore literary speech representation. *The Routledge handbook of corpus linguistics*, 531-544.

Anderson, J., Beavan, D., & Kay, C. (2007). SCOTS: Scottish corpus of texts and speech. *Creating and digitizing language corpora*, 17-34. Palgrave Macmillan, London.

Anderson, W. (2013). 'Snippets of Memory': Metaphor in the SCOTS Corpus. *Language in Scotland: Corpus-Based Studies*, 215-236. Amsterdam: Brill Rodopi.

Anthony, L. (2015). Laurence Anthony's AntConc. *AntCont Software.[cited 5 Aug 2015]. Available at: http://www. laurenceanthony. net/software/antconc*.

Aston, G. (2015). Learning phraseology from speech corpora. In: Leńko-Szymańska, A.; Boulton, Alex. (Eds.). *Multiple affordances of language corpora for data-driven learning,* 65-84. Amsterdam: John Benjamins.

Bernardini, S. (2004). Corpora in the classroom: An overview and some reflections on future developments. In J.McH. Sinclair (Ed.), *How to use corpora in language teaching*, *12*, 15-36.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Boulton, A. (2008) Looking for empirical evidence of data-driven learning at lower levels. In Lewandowska-Tomaszczyk, B. (ed.), *Corpus linguistics, computer tools, and applications: State of the art*, 581–598. Frankfurt: Peter Lang.

Boulton, A. (2009a). Data-driven learning: Reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics.*, *35*(1), 81-106.

Boulton, A. (2009b). Testing the limits of data-driven learning: Language proficiency and training. *ReCALL*, *21*(1), 37-54.

Boulton, A. (2011). Data-driven learning: the perpetual enigma. In: S. Goźdź-Roszkowski (Ed.), *Explorations across Languages and Corpora*, 563-580. Peter Lang.

Boulton, A. (2012) Hands-on / hands-off: Alternative approaches to data-driven learning. In: Thomas, J. and Boulton, A. (Eds.), *Input, process and product: Developments in teaching and language corpora*, 152–168. Brno: Masaryk University Press.

Boulton, A., & Tyne, H. (2013). Corpus linguistics and data-driven learning: a critical overview. *Bulletin suisse de linguistique appliquée*, 97, 97-118

Boulton, A. (2017). Research timeline: Corpora in language teaching and learning. *Language Teaching*, 50(4): 483-506.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, *67*(2), 348-393.

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, *17*(1), 47-64.

Braun, S. (2006). ELISA–a pedagogically enriched corpus for language learning purposes. In: S. Braun, K. Kohn & J. Mukherjee (Eds.). *Corpus technology and Language Pedagogy: New Resources, New Tools, New Methods*, 25-47. Frankfurt/M: Peter Lang.

Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL*, *19*, 307-328.

Brezina, V., Love, R., & Aijmer, K. (2018). Corpus Linguistics and Sociolinguistics: Introducing the Spoken BNC2014. *Corpus Approaches to Contemporary British Speech*, 3-9. Routledge.

Brumfit, C. J., & Carter, R. (1986). *Literature and language teaching*. Oxford University.

Breyer, Y.A. (2009). Learning and teaching with corpora: Reflections by student teachers. *Computer Assisted Language Learning*, *22*(2), 153-172.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 711-733.

Carr, P. (1999). *English phonetics and phonology: An introduction*. Blackwell.

Chambers, A. (2010). What is data-driven learning?. *The Routledge handbook of corpus linguistics*, 345-358.

Chela-Flores, B. (2003). Optimizing the teaching of English suprasegmentals. *Bells: Barcelona English language and literature studies*, *12*.

Chela-Flores, B. (1997). Rhythmic patterns as basic units in pronunciation teaching. *Onomazein*, *2*, 111-134.

Chujo, K., & Oghigian, K. (2008). A DDL approach to learning noun and verb phrases in the beginner level EFL classroom. *Proceedings of TaLC*, 65-71.

Cook, G. (1994). *Discourse and literature: The interplay of form and mind*. Oxford: Oxford University Press.

Debrock, M., Flament-Boistrancourt, D., & Gevaert, R. (1999). Le manque de "naturel" des interactions verbales du non-francophone en français. Analyse de quelques aspects à partir du corpus LANCOM. *Faits de langues*, *13*(1), 46-56.

Field, J. (2003). Promoting perception: Lexical segmentation in L2 listening. *ELT journal*, *57*(4), 325-334.

Field, J. (2008). Bricks or mortar: which parts of the input does a second language listener rely on?. *TESOL quarterly*, *42*(3), 411-432.

Frankenberg-Garcia, A. (2012). Raising teachers' awareness of corpora. *Language Teaching*, *45*(4), 475-489.

Giegerich, H. J. (1992). *English phonology: An introduction*. Cambridge University Press.

Gilbert, J. B. (2008). Teaching pronunciation using the prosody pyramid. New York: Cambridge University Press

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language teaching*, *40*(2), 97-118.

Granger, S., & Gilquin, G. (2010). How can DDL be used in language teaching?. *The Routledge Handbook of Corpus Linguistics*, 359-370. London: Routledge.

Guan, X. (2013). A study on the application of data-driven learning in vocabulary teaching and leaning in China's EFL class. *Journal of Language Teaching and Research*, *4*(1), 105.

Gómez Guinovart, X. (2019). Enriching parallel corpora with multimedia and lexical semantics: From the CLUVI Corpus to WordNet and SemCor. In I. Doval & M. T. Sánchez Nieto (Eds.), *Parallel corpora for contrastive and translation studies: new resources and applications*, 141-158. Amsterdam: John Benjamins.

Hall, G. (2005). *Literature in language education*. London: Palgrave Macmillan.

Hasebe, Y. (2015). Design and Implementation of an Online Corpus of Presentation Transcripts of TED Talks. *Procedia: Social and Behavioral Sciences*, *198*(24): 174–182.

Hernández, P. S. (2011). The potential of literacy texts in the language classroom: The study of linguistic functions. *Odisea: Revista de Estudios Ingleses*, *12*, 233-244.

Hoffstaedter, P., & Kohn, K. (2009). Real language and relevant language learning activities: insights from the SACODEYL project. *The workings of the anglosphere. Contributions to the study of British and US-American cultures*, 291-303. Trier: WVT.

Jilka, M. (2000). Testing the contribution of prosody to the perception of foreign accent. *New Sounds*, *4*, 199-207.

Johns, T. (1991a). Should you be persuaded: Two Samples of Data-Driven Learning. In T. Johns & P. King (Eds.), *Classroom Concordancing*. *English Language Research Journal*, *4*, 1-16

Johns, T. (1991b). From printout to handout: Grammar and vocabulary teaching in the context of datadriven learning. In T. Johns & P. King (Eds.), *Classroom Concordancing*. *English Language Research Journal*, *4*, 27-45.

Johns, T. (1997). Contexts: The background, development and trialling of a concordance-based CALL program. *Teaching and language corpora*, 100-115.

Johns, T. F., Hsingchin, L., & Lixun, W. (2008). Integrating corpus-based CALL programs in teaching English through children's literature. *Computer Assisted Language Learning*, *21*(5), 483-506.

Jones, R. H. (1997). Beyond "listen and repeat": Pronunciation teaching materials and theories of second language acquisition. *System*, *25*(1), 103-112.

Kasper, L. (2000). New technologies, new literacies: Focus discipline research and ESL learning communities. *Language Learning & Technology*, *4*(2): 96–116.

Kennedy, G. (1992). Preferred ways of putting things with implications for language teaching. *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, 335-373.

Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language learning & technology*, *5*(3), 77-90.

Kim, Y. S. (2007). A Zero-Relative and Intonational Phrase Boundary. *영어학*, *7*(4), 587-611.

Kjellin, O. (1999). Accent Addition: Prosody and Perception Facilitates Second Language Learning. In O. Fujimura, B. D. Joseph, & B. Palek (Eds.), *Proceedings of LP'98 (Linguistics and Phonetics Conference) at Ohio State University, Columbus, Ohio, September 1998*, *2*, 373- 398. Prague: The Karolinum Press.

Kreidler, C. W. (1993). *The pronunciation of English: a course book in Phonology*. Blackwell.

Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and Thought* (2nd ed.). Cambridge University Press

Lang, M. & Gómez Guinovart, X. (2021). Developing and Implementing an English-Spanish Literary Parallel Audio-textual Corpus for Data-driven ESL Learning. *DELTA, Documentação e Estudos em Linguística Teórica e Aplicada* 37(1).

Lee, S. (2011). Challenges of using corpora in language teaching and learning: Implications for secondary education. *Linguistic Research*, *28*(1), 159-178.

Leech, G. (1992). Corpora and theories of linguistic performance. *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, 105-122.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2007). The Talkbank Project. *Creating and Digitizing Language Corpra*, 163-180.

Mauranen, A. (2004). Spoken corpus for an ordinary learner. *How to use corpora in language teaching*, *12*, 89.

McCarthy, M. (2008). Accessing and interpreting corpus information in the teacher education context. *Language Teaching*, *41*(4), 563-574.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Routledge.

McEnery, T., & Xiao, R. (2011). What corpora can offer in language teaching and learning. *Handbook of research in second language teaching and learning*, *2*, 364-380.

McIntyre, D., & Walker, B. (2019). *Corpus Stylistics: Theory and Practice*. Edinburgh University Press.

McIntyre, D. (2015). Towards an integrated corpus stylistics. *Topics in Linguistics*, *16*(1), 59-68.

McKay, S. (1982). Literature in the ESL classroom. *Tesol Quarterly*, *16*(4), 529-536.

Mishan, F. (2004). Authenticating corpora for language learning: a problem and its resolution. *ELT journal*, *58*(3), 219-227.

Mukherjee, J. (2004). Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany. *Applied Corpus Linguistics*, 239-250. Brill Rodopi.

Mukherjee, J. (2006). Corpus linguistics and language pedagogy: The state of the art–and beyond. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 5-24.

Mukherjee, J., & Rohrbach, J. M. (2006). *Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research*, 205-232.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. *How to use corpora in language teaching*, *12*, 125-156.

Pennington, M. C. (1996) Phonology in English Language Teaching: An International Approach. Longman.

Römer, U. (2008). Corpora and language teaching. *Corpus linguistics. An international handbook*, *1*, 112-130.

Römer, U. (2011). Corpus research applications in second language teaching. *Annual review of applied linguistics*, *31*, 205-225.

Schlüter, J. (2009). *Rhythmic grammar: The influence of rhythm on grammatical variation and change in English*. Walter de Gruyter.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied linguistics*, *11*(2), 129-158.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. *Attention and awareness in foreign language learning*, *9*, 1-63.

Schmidt, R. W. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction*, 3-32. New York: Cambridge University Press

Schmidt, T. (2014, May). The research and teaching corpus of spoken German — FOLK. *LREC*, 383-387.

Shepherd, T. M., & Sardinha, T. B. (2013). A rough guide to doing corpus stylistics. *Matraga-Revista do Programa de Pós-Graduação em Letras da UERJ*, *20*(32), 66-89.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, J. (2005). Corpus and text — basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice*, 1-16. Oxford: Oxbow Books.

Smart, J. (2014). The role of guided induction in paper-based data-driven learning. *ReCALL: the Journal of EUROCALL*, *26*(2), 184.

Solé Sabater, M. J. (1991). Stress and rhythm in English. *Revista alicantina de estudios ingleses*, *4*, 145-162.

Sotelo Dios, P., & Gómez Guinovart, X. (2012). A multimedia parallel corpus of English-Galician film subtitling. *1st Symposium on Languages, Applications and Technologies*. Schloss Dagstuhl — Leibniz-Zentrum für Informatik.

Sotelo Dios, P. (2016). Adquisición de competencias en traducción audiovisual mediante un corpus multimedia. In D. Gallego Hernández (Ed.), *New insights into corpora and translation*,1-16. Cambridge Scholars Publishing.

Sripicharn, P. (2010). How can we prepare learners for using language corpora. *The Routledge handbook of corpus linguistics*, 371-384.

Stockwell, R. P., & Bowen, J. D. (1965). *The Sounds of English and Spanish*. Chicago: University of Chicago Press.

Talai, T., & Fotovatnia, Z. (2012). Data-driven learning: A student-centered technique for language learning. *Theory and Practice in Language Studies*, *2*(7), 1526.

Tan, P. K. (2013). Literary discourse. *The Routledge Handbook of Discourse Analysis*, 654-667. Routledge.

Tannen, D. (1982). Oral and literate strategies in spoken and written narratives. *Language*, 1-21.

Thompson, P., & Tribble, C. (2001). Looking at citations: Using corpora in English for academic purposes. *Language Learning & Technology*, *5*(3), 91-105.

Trofimovich, P., & Baker, W. (2007). Learning prosody and fluency characteristics of second language speech: The effect of experience on child learners' acquisition of five suprasegmentals. *Applied Psycholinguistics*, *28*(2), 251-276.

Tyne, H. (2012). Corpus work with ordinary teachers: Data-driven learning activities. *Input, Process & Product: Developments in Teaching and Language Corpora*, 114-129. Masaryk University Press.

Vyatkina, N. (2016a). Data-driven learning for beginners: The case of German verb-preposition collocations. *ReCALL*, *28*(2), 207–226.

Vyatkina, N. (2016b). Data-driven learning of collocations: Learner performance, proficiency, and perceptions*. Language Learning & Technology*, *20*(3), 159–179.

Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and cognitive processes*, *19*(6), 713-755.

Whitley, M. S. (2002). *Spanish/English contrasts: A course in Spanish linguistics*. Georgetown University Press.

Widdowson, H. G. (1978). *Teaching language as communication*. Oxford University Press.

Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied linguistics*, *21*(1), 3-25.

Wieser, J. (2005). The comprehension and acquisition of metaphorical and idiomatic phrases in L2 English. *Linguistics, Language Learning and Language Teaching*, *10*, 131.

Wynne, M. (2006). *Stylistics: corpus approaches*. Oxford University Press.

Yoon, C. (2011). Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes*, *10*(3), 130-139.