http://dx.doi.org/10.1590/1678-460X2021370106

Articles

Developing and implementing an English-Spanish literary parallel audio-textual corpus for data-driven ESL learning

Desenvolvimento e implementação de um corpus áudio-textual paralelo literário inglês-espanhol para a aprendizagem baseada em dados de ESL

Michael Lang[1]
Xavier Gómez Guinovart[2]

ABSTRACT

The purpose of this paper is to present the LITTERA corpus, an English-Spanish literary parallel speech corpus created for the purpose of language learning, and to sketch out a few pedagogical applications for the study of English phonology by Spanish-speaking language learners. It is composed of 25 literary texts that have been aligned with the Spanish translation and are accompanied by audio from the corresponding audiobooks. In this article, we will detail its conception, composition and features at length, as well as provide a few examples of how LITTERA can be applied in language learning, particularly within the realm of oral comprehension and speech production.

Keywords: *Data-Driven Learning (DDL); parallel corpus; audio-textual corpus; Second Language Acquisition (SLA).*

[1] Universidade de Santiago de Compostela. Santiago de Compostela – Espanha. http://orcid.org/0000-0003-3268-215X. E-mail: mjlang827@gmail.com.

[2] SLI/TALG - Universidade de Vigo. Vigo, Pontevedra, Espanha. http://orcid.org/0000-0001-9961-6953. E-mail: xgg@uvigo.gal.

RESUMO

O objetivo deste trabalho é apresentar o corpus LITTERA, um corpus áudio-textual literário inglês-espanhol criado para a aprendizagem de línguas, e esboçar algumas aplicações pedagógicas para o estudo da fonologia inglesa por estudantes de língua espanhola. O corpus LITTERA é composto por 25 textos literários que foram alinhados com a sua tradução espanhola e são acompanhados pelo áudio dos audiolivros correspondentes. Neste artigo, vamos detalhar a sua concepção, a sua composição e as suas características em profundidade, bem como fornecer alguns exemplos de como o corpus LITTERA pode ser aplicado na aprendizagem de línguas, particularmente nos domínios da compreensão oral e da produção de fala.

Palavras-chave: *Aprendizagem baseada en dados; corpora paralelos; corpora áudio-textuais; aquisição de segunda língua.*

1. Introduction

While speech corpora are certainly nothing new to the field of Corpus Linguistics, they are rather scarce when it comes to research in Data-Driven Learning (DDL)[3]. DDL is a method of language learning in which the learner interacts directly with corpus data, either through printouts of corpus extracts or (more commonly nowadays) via an electronic interface, which often allows full access to the corpus. DDL was pioneered by Tim Johns (1991a, 1991b) and the general idea can be summed up in the following:

> "DDL typically involves exposing learners to large quantities of authentic data — the electronic corpus — so that they can play an active role in exploring the language and detecting patterns in it. They are at the centre of the process, taking increased responsibility for their own learning rather than being taught rules in a more passive mode." (Boulton 2009)

[3] This research has been carried out thanks to the project DeepReading (RTI2018-096846-B-C21) supported by the Ministry of Science, Innovation and Universities of the Spanish Government and the European Fund for Regional Development (MCIU/AEI/FEDER).

DDL has been credited with providing many benefits to learners, such as improvements in vocabulary (Römer 2008), recognition of lexico-grammatical patters (Sripicharn 2010; Smart 2014), learner autonomy, language awareness (Talai & Fotovatnia 2012) and noticing (Boulton 2011). Furthermore, DDL allows learners to take an active role in the development of their language skills through *inductive learning*. Unlike the more traditional approach of teaching the rule first and then analyzing examples and doing exercises that fit neatly within said rule, inductive learning first examines the language data and then forms rules based on the observations (Smart 2014). And because DDL provides real examples of language use, students can gain a more nuanced and well-rounded idea of how a word or expression may be used and in what contexts.

As the number of empirical studies in the field continues to grow, a fairly clear picture of the perceptions from both students and teachers has begun to emerge. Yoon's (2011) overview of the research shows that students enjoy the fact that the language in corpora is authentic and that there is context for the words or structures in question. They claim to use the corpus as a reference tool, as one would a dictionary or online automatic translator, and also report greater feelings of autonomy and confidence after having engaged in DDL activities. Such positive feedback was not without its criticisms, as some students found DDL tasks to be too time consuming for what they were getting out of them and felt frustrated when it came to understanding some of the concordance examples and formulating queries. As for teachers, an overwhelming majority goes through the training and certificate programs without ever encountering corpora, so it is no surprise that they often appear hesitant to fully embrace the idea of bringing corpora into the classroom, even after receiving introductory training, despite recognizing their pedagogical value (Mukherjee 2004; Breyer 2009). Their enthusiasm is often accompanied by skepticism regarding the more pragmatic aspects of their day-to-day praxis along with the feasibility of working corpora into an already loaded curriculum (Tyne 2012). On top of that, some teachers worry that corpora will undermine their authority and ultimately replace them, which is likely why they tend to see corpora most useful for teacher-centered activities (Mukherjee 2004). However, introducing corpora into the classroom does not make the teacher's role redundant, but rather shifts it from that of rule imparter to one of task facilitator, what Smart (2014: 187) calls *guided induction*, making sure the students are able to discover the language patterns and do not subsequently arrive at erroneous conclusions.

There is no doubt, then, that training is an important factor in DDL if students and teachers are to feel comfortable using corpora. Researchers have suggested ways to ease new users into corpora. Some (Mukherjee, 2006; McCarthy, 2008; Frankenberg-Garcia, 2012) suggest beginning with the nature of corpora by making them aware of such factors as availability, size, language(s), design, data type and domain. After understanding the basic features of corpora, Sripicharn (2010) suggests trying short strings of letters or words, and then adding or removing characters to see how the corpus treats different search queries. While becoming familiarized with a corpus may seem like a challenging task, Boulton (2009) points out "[c]orpora with integrated interfaces for on-line access today may require as little as five minutes' introduction". As will be seen in section 2.3 below, LITTERA provides a set of introductory video tutorials aimed at familiarizing the student with its search capabilities.

Since its inception, the field has grown considerably and continues to be an active area of research (Stevens 1995; Kaltenböck and Mehlmauer-Larcher 2005; Johns et al. 2008; Chang and Sun 2009; Römer 2011; Boulton 2017). However, almost all work in DDL has focused on textual corpora, usually monolingual. This is understandable given the amount of time, effort and resources that must go into a reasonably sized speech corpus, not to mention technological restrictions, especially in the field's early days. Only very recently has research begun to appear using speech corpora in language learning, the most notable for our purposes being the creation of the TED corpus[4] (Hasebe, 2015), which provides online access to audio and video from speeches at TED events and which has been used to study phraseological items in English (Aston, 2015). In response to the scarcity of speech corpora in DDL, we have created the LITTERA corpus, an English-Spanish literary parallel speech corpus, which we will present in this paper, along with some examples of its applications in the field of language learning, particularly within the study of English phonology.

[4] https://yohasebe.com/tcse/

2. The LITTERA corpus

The LITTERA corpus[5] was conceived as a tool for Spanish university students learning English. It currently exists as a sub-corpus within the CLUVI corpus[6] from the University of Vigo (Gómez Guinovart, 2019).

LITTERA is comprised of 25 fictional literary texts in English that have been aligned with their Spanish translations at the sentence level, although some translation units (TUs) may contain multiple sentences for reasons which will become clear below. Furthermore, audio from the corresponding English audiobooks has been edited into individual audio files and aligned with each corresponding TU. We can therefore describe the LITTERA corpus as a *literary parallel audio-textual corpus*.

The corpus contains nearly 2 million words (1,968,609 total: 982,115 in English; 986,494 in Spanish) and 63,693 TUs, which means there are also 63,693 individual audio files, one for every TU. The literary works span three centuries (19th, 20th and 21st) and include a variety of genres such as the novel (the most prevalent by far), short stories, young adult fiction and children's literature[7].

2.1. *Literature and language learning*

The use of literary texts is not only advantageous from the standpoint of creating a speech corpus thanks to the availability of audiobooks, but also in terms of the relevancy for students (Kasper, 2000; Johns et al., 2008). While it is not the aim of this paper to go into the numerous benefits of literature in language pedagogy, it is worth briefly making a few comments on research that has been done in this area and how it pertains to LITTERA. As Sánchez Hernández (2011) summarizes:

[5] http://sli.uvigo.gal/CLUVI/LITTERA/
[6] http://sli.uvigo.gal/CLUVI/
[7] See the LITTERA home page (http://sli.uvigo.gal/CLUVI/LITTERA/) for a detailed list of the works included in the corpus.

> “[L]iterary 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.”

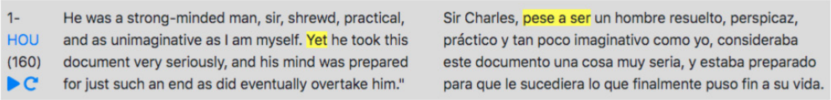
We believe that LITTERA can provide a path to both goals described by Sánchez Hernández, developing literary comprehension and sensibility and enhancing the communicative skills of language. With regards to the former, thanks to the existence of parallel texts, LITTERA may motivate students to take on literary works that may otherwise seem impenetrable, especially those that use more archaic language, e.g. Jane Austen’s *Sense and Sensibility*, but are relevant to their studies in English philology. Students can consult the translations, as well as other examples of any words or expressions, both within that specific work or in the corpus as a whole, thus allowing the students to move beyond any obstacles at the surface forms in order to develop greater literary competencies. By using an electronic corpus such as LITTERA, it is easy to get an idea of the frequency of certain difficult words or expressions in a text and how relevant those will be as they read. If, with one search query, they can see how often a word will occur in the text, that will most likely help them determine how much attention that item should receive. As for the latter objective of enhancing communicative skills, LITTERA may prove to be particularly helpful in this regard, not only in reading and writing, but also in speech production and oral comprehension, as we will see in more detail below.

Finally, it is worth commenting on the idea of *literary language* and the misguided perception of many teachers and students that the language in literary texts is of little interest to language learners or beyond their grasp. As Hall (2005) notes, “if the language of literature is in any way distinct, as has been argued, it is distinct for such a toleration of a greater variety than is found in any other kind of language use. It can include spoken and written features, diverse levels of formality, social, professional styles, dialects, sociolects and idiolects: a range of the language necessarily of interest – if undoubtedly challenging – to the language student." Thanks to this wide variety of registers and linguistic features found in most literary works as well as the variety of genres and epochs represented in the corpus, we believe that LITTERA will be of much use to the language learner, not only in terms of the written language, but also what this means for language learning through a speech corpus composed of audiobooks.

2.2. Methodology

The texts were primarily selected based on the English language curriculum at the University of Santiago de Compostela, although text selection was also contingent on whether or not both the book's Spanish translation and English audiobook were available. The translations and audiobooks had to be vetted for quality since the audiobooks were all taken from public sources and the chosen translations were those available online. Whenever possible, professionally recorded audiobooks were selected over amateur versions. Amateur versions were discarded for poor audio quality, pronunciation errors or choppy prosody, along with those whose narrators were non-native speakers of English. The selection of translations depended not only on the quality of the translation, but also alignment feasibility. This is mostly due to the nature of literary translation and its tendency to eschew very direct and mirrored translations in favor of serving the narrative and capturing the author's style, ensuring a naturalness in the L2 that very literal translations would inhibit and therefore allowing the translation to feel much more like an original text. As a result, paragraph structure is frequently altered to the point where the parallel texts are unmanageable (compared, for instance, to the proceedings from the European Union whose translations are very direct and much more 'literal'). This vetting process for both the audiobooks and the translations was fairly subjective and ultimately left up to the decision of the compiler. In the cases where the translator was unknown, it usually involved reading through sections of the text and ensuring that the translation quality was adequate and therefore conducive to learning.

The texts (once converted to plain text files) were aligned using the free and open-source LF Aligner[8] and then revised manually, once before editing the audio to ensure overall alignment quality, then again during the audio editing process since in some cases the prosody of the narration required a change in the segmentation and to sometimes place two or more sentences in a single TU. In other cases, the nature of the translation led us to keep multiple sentences in one segment, such as with the following example, where two separate sentences in English were translated into a single sentence that could not be split up due to the difference in syntax (see Figure 1).



1- HOU (160) ▶ C	He was a strong-minded man, sir, shrewd, practical, and as unimaginative as I am myself. Yet he took this document very seriously, and his mind was prepared for just such an end as did eventually overtake him."	Sir Charles, pese a ser un hombre resuelto, perspicaz, práctico y tan poco imaginativo como yo, consideraba este documento una cosa muy seria, y estaba preparado para que le sucediera lo que finalmente puso fin a su vida.

Figure 1 – Example of *sentence fusion* in the LITTERA corpus (from Arthur Conan Doyle's *The Hound of the Baskervilles*).

The audio was edited using Audacity[9] open-source sound editing software. For the most part, this was done manually, although it was possible with certain recordings to have the program automatically slice it at pauses of a specific duration. This depended on how the text was narrated and whether the pauses between sentences were significantly longer than those within a sentence, such as with commas. This method was not error-proof and would still often lead to editing problems in which the audio did not line up with the text or a sentence with a long pause in it would be split into two. Despite the laboriousness of such a task, manually editing the audio has seemingly led to a very high quality in the alignment of the texts and the audio[10].

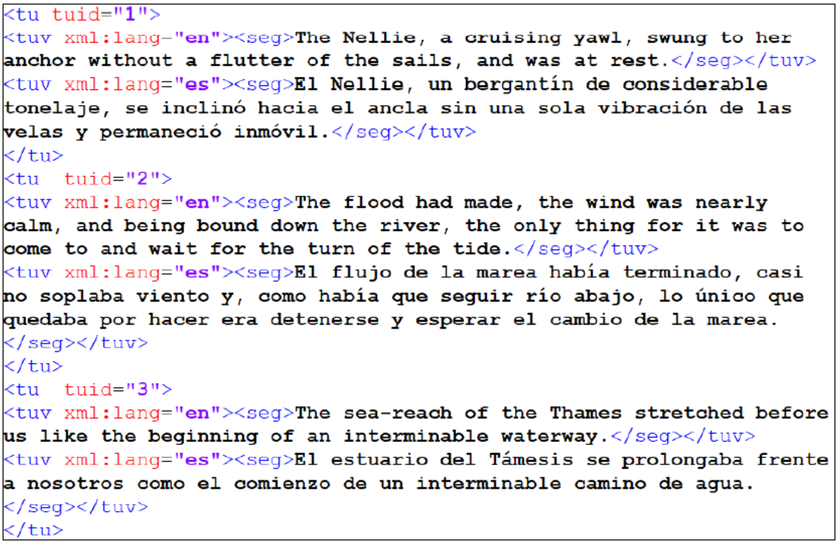
Once the texts were aligned and the audio properly sliced, the parallel text files were converted to XML format. The format chosen for storing the aligned parallel texts in CLUVI is an adaptation of the

[8] https://sourceforge.net/projects/aligner/

[9] https://www.audacityteam.org/

[10] However, it is important to note that, due to lack of human resources, we have not yet been able to obtain any measure of inter-analysis agreement to assess the reliability of the manual audio-to-text alignment.

TMX format (Savourel, 2005), as this is the XML encoding standard for translation memories, regardless of the application used. The XML specification for LITTERA encoding follows the general conventions of the TMX format, as shown in the fragment of the corpus in Figure 2.



```
<tu tuid="1">
<tuv xml:lang-"en"><seg>The Nellie, a cruising yawl, swung to her
anchor without a flutter of the sails, and was at rest.</seg></tuv>
<tuv xml:lang="es"><seg>El Nellie, un bergantín de considerable
tonelaje, se inclinó hacia el ancla sin una sola vibración de las
velas y permaneció inmóvil.</seg></tuv>
</tu>
<tu  tuid="2">
<tuv xml:lang="en"><seg>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.</seg></tuv>
<tuv xml:lang="es"><seg>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.
</seg></tuv>
</tu>
<tu  tuid="3">
<tuv xml:lang="en"><seg>The sea-reach of the Thames stretched before
us like the beginning of an interminable waterway.</seg></tuv>
<tuv xml:lang="es"><seg>El estuario del Támesis se prolongaba frente
a nosotros como el comienzo de un interminable camino de agua.
</seg></tuv>
</tu>
```

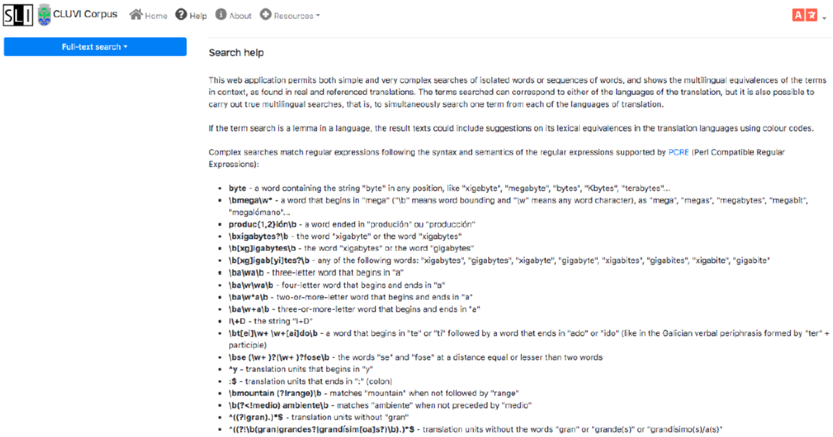
Figure 2 – Fragment of the LITTERA corpus (from Joseph Conrad's *Heart of Darkness*).

LITTERA TMX files are stored with a file name according to their title, and include the sequential number of each translation unit as an attribute *tuid* (translation unit identifier) of the *tu* (translation unit) element. Along with the text of each translation unit, the audio files are stored as audio OGG files with a file name according to their title and their sequential number. In such a way, whenever users search the LITTERA corpus they get both the bilingual text pairs and the English audio files corresponding to the text.

2.3. Search queries in LITTERA

Searches may be carried out in English, Spanish or both simultaneously and can operate using regular expressions, which allow for a more refined search. Complex searches allowed in the

CLUVI corpus match regular expressions following the formalism of the regular expressions supported by PCRE (Perl Compatible Regular Expressions)[11]. Users can find a detailed explanation of how to use the regular expressions available by clicking on the *Help* icon at the top of the page (see Figure 3). We will see some examples of how regular expressions can be used to search the corpus in the next section. Figure 4 below shows a screenshot of the search menu.



Search help

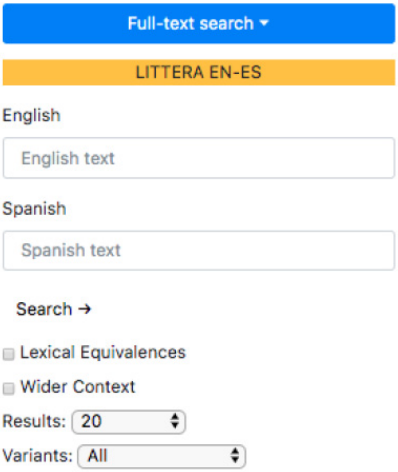
This web application permits both simple and very complex searches of isolated words or sequences of words, and shows the multilingual equivalences of the terms in context, as found in real and referenced translations. The terms searched can correspond to either of the languages of the translation, but 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)"

Figure 3 – Screenshot of LITTERA *Help* page explaining regular expressions



□ Lexical Equivalences

□ Wider Context

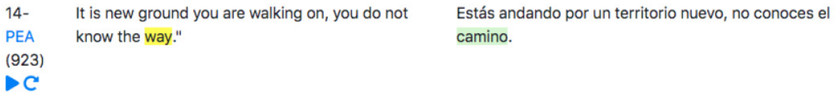
Results: 20

Variants: All

Figure 4 – Screenshot of LITTERA search menu.

[11] http://pcre.org/

Advanced options, which provide additional information in the search results, are available underneath the text boxes. The *Lexical Equivalences* option uses the WordNet-based Galnet multilingual dictionary[12] (Gómez Guinovart and Solla Portela, 2018) to highlight the cross-linguistic correspondence of the word in the translation (see Figure 5 below). If the word is found in Galnet, all the possible translations will appear at the top of the page above the search results, along with links to the entries of each use in the dictionary, as can be seen in Figure 6.

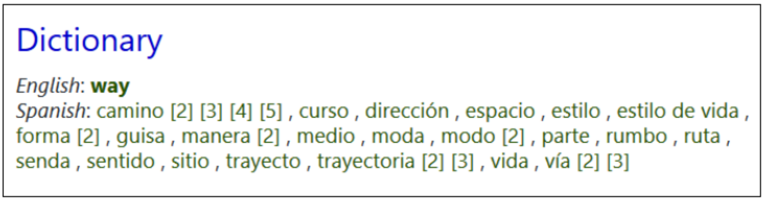


14-
PEA
(923)

It is new ground you are walking on, you do not know the way."

Estás andando por un territorio nuevo, no conoces el camino.

Figure 5 – Highlighting of lexical equivalences in LITTERA search results (from John Steinbeck ‘s *The Pearl*).



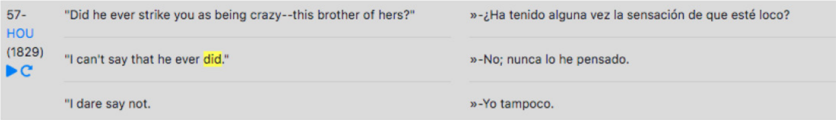
Dictionary

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]

Figure 6 – List of translations for *way* from Galnet in LITTERA search results.

The *Wider Context* option displays the surrounding translation units, which allow the user to extract more meaning from those segments that may appear ambiguous without further context. The example below (see Figure 7) shows how context can be essential to understanding the translations of certain sentences. Said in isolation, “I can’t say that he ever did” would unlikely be translated as “No; nunca lo he pensado” (lit: *No, I have never thought about it*). What is clear is that both the English and Spanish sentences beg for more context. We do not know who *he* is in the original sentence or what *he* did, and there is no mention of another person in the translation. However, with more context, the original and its translation make perfect sense.

[12] http://sli.uvigo.gal/galnet/



57- HOU (1829) ▶ C	"Did he ever strike you as being crazy--this brother of hers?"	»-¿Ha tenido alguna vez la sensación de que esté loco?
	"I can't say that he ever did."	»-No; nunca lo he pensado.
	"I dare say not.	»-Yo tampoco.

Figure 7 – Extending the segment context in LITTERA search results (from Arthur Conan Doyle's *The Hound of the Baskervilles*).

The *Results* option allows the user to limit the number of results the corpus returns (see Figure 8). Of special note is the option to limit the results to 1, 5, 10 or 15 per text, which allows the user to interact with a more manageable return size for searches that would yield hundreds or thousands of results.

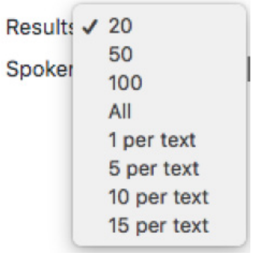
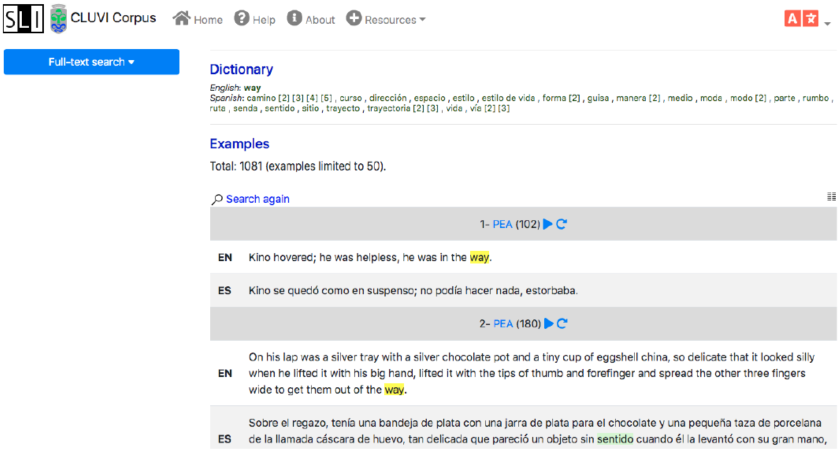


Figure 8 – Selecting number of results in LITTERA search interface.

Lastly, the *Variants* option offers the choice to search by variety, North American English or British English. The American variety contains 539,557 words and more than doubles the British variety which contains 257,580 words. Two works were excluded from this option, *Sense and Sensibility* and *Dubliners*, due to the fact they contain multiple narrators from both varieties.

Once the user has carried out a search, there are still a couple of further options available. The first is the ability to rewind the audio by 2-second intervals, the icon for which can be seen next to the play button in the images above (for instance, in Figure 7). The other is more of a formatting preference which allows the parallel texts to be viewed vertically side by side or horizontally one on top of the other. The vertical side by side view can be seen in Figures 1, 5 and 7 above while the horizontal stacked view can be seen in Figure 9 below.



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]

Examples

Total: 1081 (examples limited to 50).

Search again

1- PEA (102)

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

ES 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.

ES Sobre el regazo, tenía una bandeja de plata con una jarra de plata para el chocolate y una pequeña taza de porcelana de la llamada cáscara de huevo, tan delicada que pareció un objeto sin sentido cuando él la levantó con su gran mano,

Figure 9 – Screenshot of LITTERA search interface with stacked view in a search for *way*.

The corpus also provides full access to the full individual audiobooks with the TUs in the sequential order, the option for which can be found in the *Full Text Search* menu under *Audiobooks*. Therefore, the aligned texts can serve as reading guides and allow the user to search within a specific work, which is ideal for carrying out a more in-depth literary analysis. By searching for words or expressions and considering their frequency, students can analyze *how* an author uses language and what that means within the context of the narrative.

Finally, before moving on to the pedagogical applications of LITTERA, it should be mentioned that new users ought to take some time to familiarize themselves with the corpus and its search features, as is recommended when undertaking DDL activities more generally (Braun 2005, 2006). A series of three video tutorials is currently being developed, which will introduce the user to the corpus and its search capabilities through a series of practical exercises. Until then, users can consult the Help page and experiment with regular expressions in their own "practice" queries before taking on more specific research tasks, especially those outlined below.

3. Pedagogical applications of the LITTERA corpus

As mentioned in the introduction, most of the research that has been carried out in DDL, over the last few decades, has looked at the pedagogical applications of strictly textual corpora, therefore leaving an open space for research in how phonology can be studied within the DDL framework. While LITTERA could also be used to explore lexico-grammatical issues in English (or Spanish, for that matter), in this section we will focus on the pedagogical applications of LITTERA as they pertain to speech production and comprehension, as it is our contention that this area has been understudied in the field of DDL thus far.

When learning and practicing pronunciation, it is easy for students of English to want to focus on citation forms of words, or what Brown (1990) refers to as their *ideal* forms, rather than on how they often occur in connected speech. As Lengeris (2012) notes, "pronunciation accuracy in a second language (L2) requires mastering production of both segmental (i.e. consonants and vowels) and suprasegmental or prosodic features of speech (i.e. features that extend over more than one segment such as lexical stress, pitch accent, rhythm and intonation), but teaching pronunciation of the latter is traditionally neglected in language classrooms." Lengeris goes on to say that "perceptual training can improve not only the perception of L2 segmentals and suprasegmentals but also their production" and concludes that "we simply need to expose the learner to multiple natural tokens of the target sounds produced by various speakers" and that "exposure to authentic and variable L1 input is therefore vital in learning".

In the following sections, we will explore how LITTERA can provide the "perceptual training" Lengeris suggests. To do so, we will look at examples of a few different segmental and suprasegmental features of English which are common sources of difficulty for Spanish-speaking learners. Section 3.1 will look at how segmentals can be explored in LITTERA, specifically the regular past tense *–ed* morpheme and word-final consonant clusters, while section 3.2 will look at suprasegmentals, where we will return to the *–ed* morpheme to see how it is affected by features of connected speech. Also, in section 3.2, we will also see how modal verbs are affected by the features of connected speech by considering examples of *can* in the corpus.

The aim of what is laid forth below is to sketch the potential of LITTERA (and perhaps other speech corpora) as a resource for the study of English phonology by language learners, Spanish-speakers in this case. We hope that this may lead to empirical studies and new investigations into how phonology can be explored by language learners through DDL.

3.1. Segmental features in LITTERA

3.1.1. The past tense morpheme

The past tense –*ed* morpheme can be phonetically expressed three ways in English, [t], [d], or [ɪd]. The general rule is that pronunciation is based on the voicing of the preceding phoneme, except when that preceding phoneme is /t/ or /d/. If the preceding phoneme is voiceless, as is /k/ in *worked*, then the past tense morpheme will assimilate to the voicing of the [k], and thus be expressed as [t]. If the preceding morpheme is voiced, as is /m/ in *named*, then the past tense morpheme will assimilate to that voicing and be expressed as a [d]. Finally, in the case of the preceding phoneme being /t/ or /d/, as in *waited* or *headed* respectively, an epenthetic vowel is introduced, /ɪ/, due to the principle of separation of like sounds (Pennington 1996), leading to [ɪd].

By the time they reach university, most Spanish students of English will have been taught these rules explicitly, although from our experience even students at this level still regularly confuse these forms, often reverting to [ɪd] as the default pronunciation due to negative transfer from Spanish since their L1 does not allow for complex consonant clusters in word-final position.

Given that stories are frequently narrated in the past tense, this morpheme is ubiquitous in the corpus. However, a general search for the past tense morpheme will most likely return an unmanageable number of results. For example, if we search for all words that end with *ed*, with a PCRE regular expression such as "*ed\b*", where "*\b*" stands for any word boundary, the query returns 27,524 results. Even if we choose to focus on the first 100 translation units, they will all be from the same book (*The Pearl* by John Steinbeck) with the same narrator.

To work around this issue we can choose to limit the search results to 1, 5, 10 or 15 per text. Not only does this allow for a variety of texts, but also a wider range of voices and speaking styles. For example, if we limit the results from this search to 1 per text, there is still a wide variety of words in the past tense, such as *awakened*, *vowed*, *wanted*, *stopped*, *jerked*, *passed*, *called*, *lived*, *looked*, *decided*, *happened* and so on.

Due to the fact that searches must be carried out using English orthography, irregular verbs and non-verbs will sometimes slip through the cracks, such as *fed*, *bed* or *naked*. However, these cases should not cause the student any real confusion since they are easily discernible as non-examples of the regular past tense morpheme. The frequency of such extraneous results is negligible.

While observations and inferences can be made from a general search for the past tense morpheme, it may be more pedagogically effective to approach the search a different way. What if the student wants to falsify the voicing rule for predicting the *–ed* morpheme? To do so, we can simply use the same query as before, but now adding a voiceless phoneme before *ed* to form individual queries, such as *ked* or *ped* (excluding *ted* for the aforementioned reasons). By searching individually, students avoid complex queries requiring regular expressions and can see the frequency of each result within the corpus, which provides useful information regarding which phonemes one should perhaps pay more attention to.

Now let's say we want *all* instances where a voiceless consonant precedes *–ed* in the same search. To do so, we can formulate a PCRE query such as "*[cfkpx]ed\b|ssed\b|[sc]hed\b*".[13] This search yields 9,070 results. Limiting the results to 1 per text, examples of the past tense morpheme in the results include *advanced*, *sun-kissed*, *picked*, *stopped*, *stretched*, *pushed*, *talked* and *liked*, among others. Of course, this search will exclude some cases of voiceless consonants before *–ed*, such as *laughed*, due to the complexity of English orthography. Also,

[13] The vertical line as a regular expression means *or*. In this query, *ss* is not included in the first set of brackets because the brackets indicate *any* character contained within and as a result the corpus will only look for one instance of each letter, thus searching for a single *s* twice. In most cases, the single *s* will result in the voiced /z/, e.g. *closed*, whereas here, we want /s/.

if we were to include *h* in the first set of brackets without specifying the preceding letter, examples of voiced phonemes preceding *–ed* such as *breathed*, *bathed*, *weighed* and *sighed* would show up in the results. While this search query may appear very intimidating at first, the use of regular expressions can be very advantageous once the student understands what each part does. The above query reads as: return any instances of any of the characters *cfkpx* before *ed* at the end of a word OR *ssed* at the end of a word OR any instances of the characters *s* or *c* before *hed* at the end of a word. It should be noted here that to the authors' knowledge, no research has shed light specifically on the practicality of regular expressions in DDL activities, although "unsophisticated queries" and "frustration in forming search strings" (Boulton 2017) have been issues in previous DDL research, which suggests that certain users may require extra guidance or training to be able to manage regular expressions effectively. The authors would like to stress the fact that the above query is meant to illustrate what *can* be done with regular expressions when carrying out a search in LITTERA. We understand rational concerns regarding the practicality of such search queries and are of the opinion that the most "efficient" search may not be the most pedagogically effective in a given learning context as it is the needs and abilities of the students that will determine how corpus-driven tasks are to be carried out.

Conversely, if we want to search for all voiced consonants before *–ed*, we could once again search by each individual combination separately. However, due to the fact that there are considerably more combinations to go through than with voiceless consonants, the student may be better off searching for all instances of a voiced phoneme before *–ed*. To do so, we could place remaining letters of the alphabet (except *t* and *d* for the aforementioned, or *s* and *h* since those cases will be specified later in the same query) in brackets before *ed*. Then we can use NOT statements (^) to avoid the double *s* as well as the voiceless combination with *h*, thus resulting in the query "*[abegijl-oqru-wyz]ed\b|[^s]sed\b|[^sc]hed\b*".[14]

[14] When a hyphen is used between letters in brackets, it means *search from one letter to another*, all inclusive. In this example, *'l-o'* means searching for *l*, *m*, *n*, and *o*.

This search yields 18,320 results. Limiting the results to 1 per text, examples of the past tense morpheme include *awakened*, *vowed*, *lived*, *destroyed*, *sharpened*, *seemed*, *changed* and *issued*, among others.

Lastly, to search for the cases in which *–ed* is pronounced as / ɪd /, we can formulate the query as "*[td]ed\b*", which returns 6,793 results. However, this is perhaps the most intuitive pronunciation for Spanish speakers and would require the least amount of attention since it is most similar to the default pronunciation by many learners due to transfer from Spanish.

After examining some of the results these searches would yield, students will quickly realize that, while the morpheme is often present, Brown's *ideal* forms are far less clear-cut in connected speech. This is mainly due to features of English rhythm, particularly *assimilation* and *elision*, which will be explored further in the following section on suprasegmentals. In order to hear more fully realized forms of the past tense morpheme in the corpus, students can search for examples of *ed* at the end of an intonational phrase, most commonly marked by some type of punctuation, such as a comma, full stop, question mark, colon, semi-colon, etc. To do so, students can simply remove '*\b*' from any of the above queries and replace it with any set punctuation marks, the most common being full stops, commas and question marks. The query "*ed[,.?]*" yields 6,037 results and invariably shows a fuller form of the past tense morpheme than is likely to appear when followed by another word (unless that word begins with a vowel), which sheds light on the large role suprasegmental factors play in determining the realization of the segments in a connected speech, as will be seen in 3.2 below. This is why a search for any segment at the end of an intonation phrase illustrates how students can hone in on more *pure* or *polished* forms of the segment in question.

3.1.2. Consonant clusters

Compared to other languages, such as Spanish, English is far less restrictive with consonant clusters, allowing for such extreme examples as CCVCCCC, as in *twelfths*. As a result, consonant clusters pose much difficulty to learners of English, especially those whose L1 is

Spanish due to the phonotactic rules of the language. Word-final VCC clusters are extremely rare in Spanish, occurring only in borrowed words or proper names, e.g. *golf* or *vals*, while the word-final VCCC cluster is phonotactically impossible. Adding to the complication is the way English speakers phonetically realize such clusters with varying levels of consistency, sometimes pronouncing the full cluster, while other times simplifying them through the processes of assimilation and elision. Knowing when clusters can be simplified is invaluable information to learners of English as it not only allows them to improve comprehension, but also to avoid unnecessarily overcomplicating their pronunciation. Students can approach their corpus search with this in mind. Below we will see how a variety of English consonant clusters can be examined in LITTERA and what insights students can gain. We will limit our examination to word-final clusters, not only because these can be difficult to perceive, but also due to the fact that, as Gilbert (2008) explains, in English "the highest priority sounds are at the end of words because they give crucial grammatical cues…[and] are usually spelled with the letters *s* or *d*". Therefore, we will consider examples of VCCC and, in some cases, VCCCC consonant clusters involving the *s* and *ed* inflections.

To begin, let us return to the past tense morpheme. As stated above, the rule frequently taught to students is that the phonetic realization of *ed* depends on the voicing (in the case of [t] and [d]) or the place of articulation (in the case of [ɪd]) of the preceding phoneme. Many of the VCCC consonant clusters formed from the addition of the past tense morpheme can be difficult for Spanish speakers, such as /-lvd/ in *involved*, /-ndʒd/ in *changed*, /-ntʃt/ in *pinched*, or /-skt/ in *asked*. Any of these can be analyzed in the corpus with a query simply based on orthography, e.g. "*lved*", although in the case of "*nged*" this will produce a small number of unwanted results such as *longed*, *hanged* or *ringed*. A simple way around this is to do a specific search for some of the frequent words in the results, such as *changed* or *arranged* since the objective is the sound itself. It may also be beneficial to students to limit their search to words at the end of an intonational phrase as we saw in the previous section in order to find examples that are less influenced by the effects of connected speech.

What becomes clear from the examples in the corpus is that the clusters /-lvd/ and /-ntʃt/ are fully realized as one would expect them to be, while /-ndʒd/ and /-skt/ show instances of unexpected pronunciations. In the case of /-ndʒd/, there is a devoicing that occurs in many of the examples, causing the middle phoneme /dʒ/ to be realized as [tʃ] or [ʃ]. It isn't until the 27th result of the search "*anged*" that we can find a case where there appears to be no devoicing in the sentence "Only that you have disarranged our little deductions" from *The Hound of the Baskervilles*. All the other examples from this same narrator in this query show devoicing of the middle affricate. Another good way for students to discern voicing, besides searching for instances at the end of intonational phrases, is to consider examples in which the following word begins with a vowel. Due to the nature of English prosody and the effect of linking, which we will see in more detail in section 3.2, a word-final consonant is often resyllabified as the initial sound of proceeding word if it begins with a vowel. In the above sentence from *The Hound of the Baskervilles*, the [d] can clearly be heard, compared to [t] in most of the other instances when preceding a vowel. Not only does this help students learn what to listen for, but also shows them that they can reduce the cluster in their own speech and even replace the affricate [tʃ] with a fricative [ʃ], easing pronunciation even further.

As for /-skt/, its unexpected pronunciation has to do with elision. Although we can find it in other words in the corpus, the most common instance is in the word *asked* due to its frequent use in narrative texts. When searched for on its own, *asked* yields 692 results. Because some narrators read more carefully than others, there are many opportunities to hear the contrast between the full cluster and the reduced cluster in which the middle consonant, in this case /k/, is omitted completely. Once students are aware that the word-final /-skt/ cluster can be reduced to [st], not only will they be able to perceive this better when listening to connected speech, but they will also be able to reproduce it with relative ease instead of trying to force the pronunciation of all three consonants in the cluster, a fairly difficult task for Spanish speakers. Furthermore, examining these clusters reinforces what the student may have already learned when exploring the pronunciation of the past tense morpheme. Both the full form and reduced form of the /-skt/ cluster still ends with a voiceless consonant, [t], after both [s] and [k]. Similarly, the devoicing of /dʒ/ in the /-ndʒd/ cluster also leads to the realization of the past tense morpheme as [t]. Finally, students have the opportunity to further discover that by searching for *asked* at the end of an intonational phrase as we saw above, one is more likely to hear the cluster in its full form.

Turning to word-final clusters involving the *s* inflection, there are a number of examples that may be confusing for Spanish speakers due to the phonotactic restrictions of Spanish, in particular /-pts/ as in *accepts*, /-kts/ as in *facts*, /-nts/ as in *accents*, /-nds/ as in *hands*, /-ksts/ as in *texts*, /-fθs/ as in *fifths*, /-fts/ as in *gifts*. As mentioned above, the most useful information that the corpus can provide for students for their own speech is discovering "easier" ways of producing these clusters. What the results in the corpus show is that in the case of /-pts/, /-kts/, /-nts/ and /-nds/ there is, invariably, a clear elision of the middle consonant. In the case of /-ksts/, there are only two results and in both cases the cluster is fully realized. However, its relatively low frequency also suggests that this cluster may not be of high priority for students. There are no examples of /-fθs/ in the corpus, although 15 instances of /-fθ/ (in *fifth*) and 3 examples /-lfθ/ (in *twelfth*) can be found, all fully realized. Finally /-fts/ can be heard in both its full form and its reduced form with an elided /t/.

Because of their complex nature for Spanish speakers, consulting multiple instances of a cluster to find out which ones can be reduced and how is of great benefit to both comprehension and production, and may also provide useful insights into English phonology more generally. The same types of queries can be formulated to examine consonant clusters in non word-final positions.

3.2. Suprasegmental Features in LITTERA

The importance of prosody for learners of English cannot be overstated since much research has shown how suprasegmental features impact intelligibility in L2 speakers (Lengeris 2012; Piske 2012) and may also lead to a more rapid improvement of segmentals (Levis and McCrocklin 2018). In fact, Kjellin (1999) notes the importance of natural fluidity in speech and suggests that "overarticulated speech has reduced intelligibility for the native listener, due to its reduced degree of coarticulation." Therefore, it is not only important for students to be aware of certain characteristics of English prosody for their own comprehension's sake, but also for the sake of being effectively understood by their interlocutors. Below we examine two examples of how LITTERA could serve as a type of *informant,* as Johns (1991a) would say, on the topic of English prosody.

3.2.1. The past tense morpheme

As noted before, students will quickly realize that despite the rules governing this morpheme being rather simple and clear, the actual presence of the past tense morpheme in fluid, connected speech can be difficult to perceive, much less distinguish in terms of the [d] vs [t] dichotomy (setting aside [ɪd] for aforementioned reasons).

Students, therefore, must take into consideration how this morpheme is affected by the suprasegmental elements at work. Where in the previous section we saw how voicing of the preceding phoneme determined the pronunciation of the *–ed* morpheme in its *ideal* form, now we can see how other factors beyond the word itself influence the morpheme's pronunciation, in this case the place of articulation of the proceeding phoneme.

Much like in previous section, we can search for all instances of the regular past tense morpheme, though this time specifying that it does NOT occur at the end of an intonational phrase by simply putting a space afterwards. What's more, we can specify which proceeding sounds we want to focus on. For example, if we want to see how the voiceless post-alveolar fricative affects the past tense morpheme, we can formulate the query "*[^td]ed sh*". This search yields a mere 114 results, but effectively illustrates the effect /ʃ/ has on the past tense morpheme. We need not look further than the very first example from *The Pearl* in which we find *bunched shadows*, which the narrator reads as /bʌn-ʃæ-doz/, eliding the past tense morpheme altogether and assimilating the preceding /ʧ/ in terms of manner.

We can carry out this same type of search with /θ/ and /ð/ by changing *sh* to *th*. Here, there is a drastic difference in frequency as this search yields 2,886 results. In order to make this a more manageable batch, we can choose to limit the results to 5 per text. If we look at the ninth result on the page, from *The Cask of Amontillado*, we will see the same verb with the past tense morpheme repeated in the same TU: “We *passed* through a range of low arches, descended, *passed* on, and descending again, arrived at a deep crypt...” The narrator elides the past tense morpheme before *through*, yet fully realizes it before *on*. What’s more, the /t/ appears to form the onset of the syllable /an/, making it [tan], a process commonly referred to as *linking* (Gilbert 2008), and more technically known as *resyllabification* (Field 2003). Thus, when said in connected speech, *passed on* appears to be [pæs-tan] due to the word-final /t/ linking the two words together in connected speech. This is a common feature of English prosody when a word-final consonant is followed by a word-initial vowel. Much like how the past tense morpheme is always present at the end of an intonational phrase, the same thing occurs before a vowel, except in these cases the consonant functions more like the onset of the next syllable than the coda of the syllable it would normally pertain to if said in isolation or at the end of an intonational phrase. This is another important feature of English prosody that would be worth students’ time to explore in the corpus.

Another way to illustrate the phenomenon of assimilation in English prosody is to explore *used to* in the corpus and compare it to *used* when followed by an object. *Used to* is frequently taught in EFL classrooms as a separate phraseological item, along with *be used to* and *get used to*. A good way to approach this in a classroom setting could be to first search the corpus for all instances of *used* when not followed by *to* or any word that begins with /t/. This can be done with a very simple query: “ *used [^t]*”. This search yields 67 results. Because of the voiced *s* [/z/] at the end of the verb *use*, the past tense morpheme is /d/ when unaffected by other prosodic elements. This is best illustrated by looking at the cases where it is followed by a vowel, e.g. *used a* or *used on*. If we want to include examples at the end of an intonational phrase, we must specify the punctuation we want to appear, as in the query “ *used[,.?!;:”][^t]*”. By leaving a space in the brackets, we do not exclude any results from the earlier search; we are simply adding more possibilities. This search yields 72 results, adding five examples of *used* at the end of an intonational phrase.

A search for “ *used to\b”* yields 180 results, all which show an invariable elision of the past tense morpheme and assimilation in voicing as the /z/ from *used* becomes /s/, even in instances where the narrator reads more slowly or more carefully. What makes this example even more illuminating is that the vowel in *to* is often reduced to a schwa when followed by a verb, as is the case with the phrase *used to think* in the 11th result from *Sense and Sensibility*: “I am sure it will all end well, and there will be no difficulties at all, to what I used to think.” This type of blending and linking is crucial to English prosody and can be studied without difficulty through the data in the LITTERA corpus.

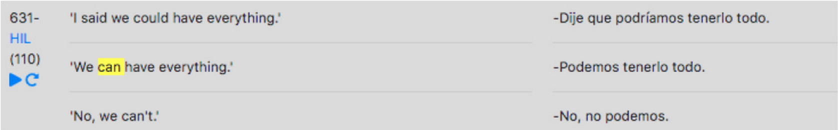
3.2.2. Function words in LITTERA: A look at *can* and *can’t*

Perhaps one of the most important aspects of English prosody is what Gilbert (2008) refers to as *contrastive highlighting/obscuring*, which she describes as “essential to the English stress and emphasis system”. This refers how English speakers use stress to highlight certain items within the sentence (usually content words) and obscure other forms that may carry very little lexical weight and contribute relatively little to the actual meaning of the sentence (usually function words). In most languages, pitch and vowel duration are the main cues by which stress is perceived. However, “in English there is an extra cue for stress which is *vowel quality,* in particular the reduced quality…of unstressed syllables” (Solé-Sabater 1991). Pennington (1996) explains this is done through “the neutralization of vowels, in most cases resulting in schwa.” Other ways of neutralizing vowels are reducing them to any of the other middle vowels, mainly /ɪ/ and /ɛ/. It’s important to keep in mind, however, that any word can be stressed in a sentence, even function words that do not normally receive stress, either for special emphasis or to contrast something said previously. Examples of this can be found in the corpus, such as the following segment from *Sense and Sensibility* in which *THEY* in all caps is stressed accordingly by the narrator: “A great deal too handsome, in my opinion, for any place THEY can ever afford to live in.”

Much like with blending and linking through assimilation and elision, vowel quality in connected speech has yet to be thoroughly examined through the lens of DDL. In this section we will take a general look at how stress affects vowel quality in English prosody by examining the modal verb *can* and its negative form *can't*.

In connected speech, the open vowel in *can*, /æ/, is regularly reduced to a middle vowel, either /ɪ/ or /ɛ/. *Can't*, on the other hand, always maintains the open vowel, which is what allows native speakers to differentiate between the affirmative and the negative in fluid speech, even if the /t/ goes unaspirated, as is frequently the case. When expressed as *cannot*, however, we can find examples of the same vowel reduction in *can* when the stress is placed on *not*, as well as the full vowel when the stress is placed on *can*. This preference does not appear to be the result of any prosodic features, but rather the personal preference of the narrator, as we will see below.

We can do a general search that would include both *can* and *can't*, but due to the disparity in frequency of the two terms, it would be much more advantageous for the student to examine them separately. To find all instances of *can*, we can use the query “\bcan\b[^’]”.[15] This search yields 1,555 results. As one would expect of a modal verb, it most often occurs before another verb, resulting in an overwhelming tendency to reduce the open vowel to a middle vowel in order to highlight the main verb of sentence. This is evident in most examples of this search. Nevertheless, we can still find certain cases where the open vowel is maintained for reasons of emphasis, as occurs in the example shown in Figure 10.



631- HIL (110) ▶ C	'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 10 – Screenshot of LITTERA search results (from Ernest Hemingway's *The Hills Like White Elephants*).

[15] A NOT statement for an apostrophe had to be included in the search despite setting word boundaries due to the fact that dialogue is often marked with apostrophes instead of quotation marks, which causes the corpus to incorrectly interpret the apostrophe in *can't* as a word boundary, unless told to do otherwise.

In this example, the narrator stresses *can* because of the contrast in the dialogue. What the two characters are debating is the *ability* to have everything; therefore, *have* is no longer the focus of the sentence. What examples like this illustrate is that even though the corpus does not contain any spontaneous speech, literary texts provide dialogue that imitates or represents what is meant to be spontaneous speech and is the closest possible approximation without being the real thing. That being said, what makes an effective narrator is his or her ability to bring out the linguistic nuances of the discourse represented in the dialogue, akin to how a theater actor gives life to a given text.

There are also some examples of *can* at the end of an intonational phrase. Here, as to be expected, the vowel is open without fail. An example of this can be found in the following segment from *The Fault in Our Stars*: "God, grant me the serenity to accept the things I cannot change, the courage to change the things I can, and the wisdom to know the difference."

If we turn to *can't*, a basic search will yield 387 results, all of which consistently provide examples of the open vowel (/æ/ or /a/). Students may even pick up on the generally unaspirated /t/ at the end of the word, which reinforces the importance of vowel quality over extremely "correct" and excessively polished pronunciation.

There are some TUs in which we can find both *can* and *can't*, as in the following taken from *Sense and Sensibility*, where the student can hear both the reduced form of *can* juxtaposed with the full vowel of *can't*: "I can tell you, and you can't think how disappointed he will be if you don't come to Cleveland."

Lastly, a search for the uncontracted form *cannot* will yield 191 results. Here, vowel quality is simply a matter of where the speaker decides to place the word stress, either on *can* or *not.* If the tonic syllable is *can*, there is no reduction, whereas vowel reduction is evident when the tonic syllable is *not*. Interestingly, students will be able to notice that even when *can* receives the tonic syllable, the vowel quality of *not* never changes.

4. Final remarks

In this paper, we introduced the LITTERA corpus, a literary parallel speech corpus composed of English language literary texts aligned with their Spanish translations and with the segmented audio from the corresponding audiobooks. We then looked at how LITTERA could be used to explore both segmental and suprasegmental features of English, in particular, the past tense morpheme, consonant clusters and the modal verb *can*. We hope the possibilities laid forth in this article may lay the groundwork for empirical research to be carried out on the study of phonology through Data-Driven Learning, whether it be with LITTERA or other pedagogically effective speech corpora. Even though LITTERA was designed with Spanish university students in mind as the main users, any student of English can benefit from the phonological data in the corpus. Likewise, LITTERA may also serve as an empirical resource in the preparation of material for English phonetics and phonology courses at the undergraduate and graduate level, as well as in providing teachers with copious examples of authentic speech from native speakers, each paired with its transcript, that can be exploited for other pedagogical purposes not discussed in the present work. We hope that LITTERA may be an important step in creating new in-roads within Data-Driven Learning given that very little research has been done on speech comprehension and production from the DDL perspective.

References

ASTON, Guy. 2015. Learning phraseology from speech corpora. In: LEŃKO-SZYMAŃSKA, Agnieszka; BOULTON, Alex. (Eds.). *Multiple affordances of language corpora for data-driven learning,* 65-84. Amsterdam: John Benjamins.

BOULTON, Alex. 2009. Data-driven learning: Reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics*, *35*(1): 81-106.

______. 2011. Data-driven learning: the perpetual enigma. In: GOŹDŹ-ROSZKOWSKI, Stanisław. (Ed.). *Explorations across Languages and Corpora,* 563-580. Frankfurt: Peter Lang.

______. 2017. Research timeline: Corpora in language teaching and learning. *Language Teaching*, 50(4): 483-506.

BRAUN, Sabine. 2005. From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, 17(1): 47-64.

______. 2006. ELISA – a pedagogically enriched corpus for language learning purposes. In: BRAUN, Sabine; KOHN, Kurt; MUKHERJEE, Joybrato. (Eds.). *Corpus technology and language pedagogy: New resources, new tools, new methods,* 25-47. Frankfurt: Peter Lang.

BREYER, Yvonne A. 2009. Learning and teaching with corpora: Reflections by student teachers. *Computer Assisted Language Learning*, *22*(2): 153-172.

BROWN, Gillian. 1990. *Listening to spoken English*. New York: Routledge.

CHANG, Wen-Li; SUN, Yu-Chih. 2009. Scaffolding and web concordancers as support for language learning. *Computer Assisted Language Learning*, *22*(4): 283-302.

FIELD, John. 2003. Promoting perception: Lexical segmentation in L2 listening. *ELT Journal*, 57(4): 325-334.

FRANKENBERG-GARCIA, Ana. 2012. Raising teachers' awareness of corpora. *Language Teaching*, *45*(4): 475-489.

GILBERT, Judy B. 2008. *Teaching pronunciation: Using the prosody pyramid*. New York: Cambridge University Press.

GÓMEZ GUINOVART, Xavier. 2019. Enriching parallel corpora with multimedia and lexical semantics: From the CLUVI Corpus to WordNet and SemCor. In: DOVAL, Irene; SÁNCHEZ NIETO, M. Teresa. (Eds.). *Parallel Corpora for Contrastive and Translation Studies: New resources and applications*, 141-158. Amsterdam: John Benjamins.

GÓMEZ GUINOVART, Xavier; SOLLA PORTELA, Miguel A. 2018. Building the Galician wordnet: Methods and applications. *Language Resources and Evaluation*, 52(1): 317-339.

HALL, Geoff. 2005. *Literature in language education*. London: Palgrave Macmillan.

HASEBE, Yoichiro. 2015. Design and implementation of an online corpus of presentation transcripts of TED Talks. *Procedia: Social and Behavioral Sciences*, 198(24): 174–182.

JOHNS, Tim. 1991a. Should you be persuaded: Two samples of data-driven learning. In: JOHNS, Tim; KING, Philip. (Eds.). *English Language Research Journal 4: Classroom Concordancing*, 1-16. Birmingham: University of Birmingham.

______. 1991b. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In: JOHNS, Tim; KING, Philip. (Eds.). *English Language Research Journal 4: Classroom Concordancing*, 27-45. Birmingham: CELS, The University of Birmingham.

JOHNS, Tim F.; HSINGCHIN, Lee; LIXUN, Wang. 2008. Integrating corpus-based CALL programs in teaching English through children's literature. *Computer Assisted Language Learning*, 21(5): 483-506.

KALTENBÖCK, Gunther; MEHLMAUER-LARCHER, Barbara. 2005. Computer corpora and the language classroom: On the potential and limitations of computer corpora in language teaching. *ReCALL*, 17(1): 65-84.

KASPER, Loretta. 2000. New technologies, new literacies: Focus discipline research and ESL learning communities. *Language Learning & Technology*, 4(2): 96-116.

KJELLIN, Olle. 1999. Accent addition: Prosody and perception facilitates second language learning. In: FUJIMURA, Osamu et al. (Eds.). *Proceedings of LP'98 (Linguistics and Phonetics Conference)*, Vol. 2, 373-398. Prague: Karolinum.

LENGERIS, Angelos. 2012. Prosody and second language teaching: Lessons from L2 speech perception and production research. In: ROMERO-TRILLO, Jesús. (Ed.). *Pragmatics and prosody in English language teaching*, 25-40. Springer, Dordrecht.

LEVIS, John M.; McCROCKLIN, Shannon. 2018. Reflective and effective teaching of pronunciation. In: ZERAATPISHE, Mitra et al. (Eds.). *Issues in Applying SLA Theories toward Reflective and Effective Teaching*, 77-89. Leiden: Brill.

McCARTHY, Michael. 2008. Accessing and interpreting corpus information in the teacher education context. *Language Teaching*, *41*(4): 563-574.

MUKHERJEE, Joybrato. 2004. Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany. In: CONNOR, Ulla; UPTON, Thomas A. (Eds.). *Applied Corpus Linguistics*, 239-250. Amsterdam: Rodopi.

______. 2006. Corpus linguistics and language pedagogy: The state of the art–and beyond. In: BRAUN, Sabine; KOHN, Kurt; MUKHERJEE, Joybrato. (Eds.). *Corpus technology and language pedagogy: New resources, new tools, new methods*, 5-24. Bern: Peter Lang.

NERI, Ambra; MICH, Ornella; GEROSA, Matteo; GIULIANI, Diego. 2008. The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, 21(5): 393-408.

PENNINGTON, Martha C. 1996. *Phonology in English Language Teaching: An International Approach*. London and New York: Longman.

PISKE, Thorsten. 2012. Factors affecting the perception and production of L2 prosody: Research results and their implications for the teaching of foreign languages. In: ROMERO-TRILLO, Jesús. (Ed.). *Pragmatics and Prosody in English Language Teaching*, 41-59. Dordrecht: Springer.

RÖMER, Ute. 2008. Corpora and language teaching. In: LÜDELING, Anke; KYTÖ, Merja. (Eds.). *Corpus linguistics. An international handbook*, *1*, 112-130. Berlin: Mouton de Gruyter.

______. 2011. Corpus research applications in second language teaching. *Annual Review of Applied Linguistics*, 31: 205-225.

SÁNCHEZ HERNÁNDEZ, Purificación. 2011. The potential of literacy texts in the language classroom. *Odisea: Revista de Estudios Ingleses*, *12*: 233-244.

SAVOUREL, Yves. 2005. *TMX 1.4b Specification*. Washington: Localisation Industry Standards Association.

SMART, Jonathan. 2014. The role of guided induction in paper-based data-driven learning. *ReCALL: the Journal of EUROCALL*, *26*(2): 184-201.

SOLÉ SABATER, Maria-Josep. 1991. Stress and rhythm in English. *Revista Alicantina de Estudios Ingleses*, 4: 145-162.

SRIPICHARN, Passapong. 2010. How can we prepare learners for using language corpora. In: O'KEEFFE, Anne; McCARTHY, Michael. (Eds.). *The Routledge handbook of corpus linguistics*, 371-384. Oxon: Routledge.

STEVENS, Vance. 1995. Concordancing with language learners: Why? When? What?. *CAELL Journal*, 6: 2-10.

TALAI, Touraj; FOTOVATNIA, Zahra. 2012. Data-driven learning: A student-centered technique for language learning. *Theory and Practice in Language Studies*, *2*(7): 1526-1531.

TYNE, Henry. 2012. Corpus work with ordinary teachers: Data-driven learning activities. In: THOMAS, James; BOULTON, Alex. (Eds.). *Input, Process and Product: Developments in Teaching and Language Corpora*, 114-129. Brno: Masaryk University.

YOON, Choongil. 2011. Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes*, 10(3): 130-139.

Recebido em: 12/12/2019
Aprovado em: 20/01/2021