AIN'T THAT SWEET
Reflections on scene indexing and annotation in the House Corpus Project[1]

DAVIDE TAIBI, IVANA MARENZI, QAZI ASIM IJAZ AHMAD

Abstract – This paper outlines the strategies, rationale and potential uses motivating the construction of the *House Corpus*, a one-million-word corpus that can be accessed by authorised users through the *MWSWeb* site (Taibi *et al*. 2015a) at http://openmws.itd.cnr.it. Part 1 illustrates the tools and techniques used to index the corpus data – transcriptions of all 177 episodes in the *House M.D.* series (original US version). In particular, it describes the commercially available *Elasticsearch* (https://www.elastic.co), used as an indexing, annotational and search tool. Part 2 explains that this is a multimedia corpus allowing viewings of different *types* of scene. The 6000-plus scenes in the corpus have been annotated in terms of their typological features: *Location type* (e.g. patient's hospital room; medical lab etc.); *Event type* (handover; differential diagnosis; precipitating medical event; patient examination etc.) and *Character Group type* (doctor/doctor; doctor/patient; doctor/caregiver; patient/caregiver etc.). The project envisages the development of various retrieval interfaces, initially *Words*, *Scenes* and *Dialogues*. This will make it possible to carry out searches in terms of *types* of scene and their distribution across the corpus without necessarily involving any other form of searching. Part 3 suggests the value of multimedia corpora in encouraging students to advance their critical discourse analysis (CDA) skills. As an example, it shows how the corpus can illustrate the priority of (inter)textual over lexicogrammatical considerations when formulating tag questions in oral discourse. Finally, the *Discussion* section argues that a typology of scenes appears to be an essential prerequisite for the construction of other types of access to the corpus data in subsequent stages of the project.

Keywords: House Corpus; indexing; scene annotation; functionality planning; CDA.

1. Introduction

This paper is a follow-up to the presentation of the preliminary phases of the *House Corpus Project* at *Clavier 17 – International Conference Representing and Redefining Specialised Knowledge*, held at the University of Bari (30

[1] Part 1 of this paper was written by Qazi Asim Ijaz Ahmad, Part 2 by Davide Taibi, and Part 3 by Ivana Marenzi. Davide Taibi and Ivana Marenzi collaborated in the writing of the remaining sections.

Doi Code: 10.1285/i9788883051531p151

November – 2 December 2017), where the research work so far undertaken was presented in summary form. The *House Corpus Project* is concerned with providing a tool for discourse analysis for university teachers and their students, in particular, those attracted by corpus-based explorations of the discourse structures presented in a contemporary US TV drama. As such, the paper explores assumptions about the goals and methods of corpus construction and classroom use of corpora, suggesting the need for greater alignment of corpus linguistics with the needs of university courses that engage with discourse analysis of contemporary English. To this end, the paper is divided into three parts: Part 1: Semantic Indexing of the *House Corpus*; Part 2: Scene management and scene level access; Part 3: Scene level access, scene annotation and discourse analysis.

One feature described at the Congress that needs to be addressed initially in this paper is its break with traditional descriptions of corpora exclusively in terms of words and word counts. Readers who expect the article to expand on the information given in the abstract – 177 episodes, (about) 1 million words – will perhaps be disappointed as the paper, but not necessarily the entire project, is concerned with the structuring of the search mechanism in terms of *scenes* rather than *words*. Compared to the term *word*, *scene* appears to be a neglected and undefined object within corpus studies despite the fact that scenes are central to the production and critical analysis of countless TV dramas. At the time of writing, a search in *Google Scholar* for the search string "*scenes in corpus linguistics"* produces no hits against twenty-three for "*words in corpus linguistics"*. Likewise, a specific search for "*word level indexing*" produces 145 hits, while "*scene level indexing*" produces just five. Four of these make no reference to corpus studies while only one, Salway (2007), mentions the search potential of *manual* scene-level indexing but, alas, only for the purposes of dismissing it as a possibility in the specific field of investigation in question, namely audio description:

> In the past archives such as that of the BBC have been for in-house use only, but the advent of the web creates the demand and opportunity to make them available for public access. A minimal requirement is to store production details such as title, director and genre with every programme and film. More useful though is shot- or scene-level indexing whereby keywords are associated with shots and scenes, enabling users to retrieve precise intervals of video data that match their queries, for example 'find me all scenes showing a woman on a horse'. Creating such indexing manually is prohibitively expensive in many cases, and the challenge of the semantic gap limits the scope for machines to generate keywords by analysis of the pixels in the video data. (Salway 2007, pp. 168-169)

While manual annotation may be inappropriate for the specific needs of audio description, we argue below that it can be beneficial in other specialised fields, such as discourse analysis, especially where it allows the functions of a corpus to be modified through supplementary 'tags' introduced by users. In this project, we aim to show that the possibilities of creating subprojects within the overall *House Corpus Project* depend on functionalities that allow such user-defined tags to be applied systematically. This, we believe, is an innovative approach to corpus studies which potentially assists teachers who wish to explore discourse in English in their university courses, in particular where this involves characterisation of the differences between spoken and written varieties.

In our experience, all too often corpora are exclusively dependent on word-based search mechanisms which become a straitjacket preventing discourse from being investigated *as* discourse. Indeed, our title *Ain't That Sweet* is an iconic representation of this, linked as it is to the detection of the intertextual features of discourse and specifically to the identification of a scene, as detailed in Part III, where Dr. House sings parts of this famous song's lyrics during a discussion of a patient's medical condition. Sensitivity to intertextual references is not something that word-based search and annotation techniques are noted for. Yet such an approach is central in explaining to students how discourse is rooted in shared culture. Exploring such cultural references assists understanding of discourse in English, which is why we suggest that, learning-wise, student engagement with annotation can be beneficial. Scene-level searching, searching, that is, for scenes that share (con)textual characteristics, is thus a first step towards constructing a corpus that facilitates the exploration of culture-related discourse features.

Our efforts to promote the scene to the status of a searchable unit are inevitably the result of teamwork. The paper is accordingly divided into three parts, with each author describing their contribution. Part 1 describes the construction of a corpus that combines scene-based indexing with traditional lemma-based indexing. Part 2 describes the basic design characteristics of an interface that, in addition to search functionalities, also supports manual scene annotation. Part 3 illustrates how all this constitutes a basis for those classroom projects that subscribe to the discourse analysis goals outlined in this paper.

2. Part 1: Semantic Indexing of the House Corpus

Although *Semantic Indexing* is never easy to define as the concept can be interpreted in many ways and is subject to re-interpretation in the wake of constant refinements and improvements in computational technique, for the purposes of the present article, and indeed the *House Corpus Project,* it may be looked upon as the process of mapping a set of metadata onto the transcripts of each episode of the *House M.D.* series. As such, it is a preliminary step in the goal of building a searchable online corpus. In itself, the task of building a set of metadata, while not requiring any understanding of the meaning or content of the individual episodes, *does* require considerable understanding and management of the characteristics of three distinct textual entities. These are:

(a) the *transcripts* of each TV episode which have been reconstructed from *source texts*;

(b) the *source texts*, *i.e.* the published *html* documents from which the transcripts have been retrieved; these are more extensive textual units as they include other types of text, most prominently various kinds of advertising;

(c) the *target texts* or *records*, *i.e.* the corpus-ready, machine-readable, searchable transcripts of each episode.

However, the transformation of source texts to target texts is not the only problem to be faced. While experienced readers immediately recognize a transcript as a transcript, closer inspection of episode transcripts (as defined above) will highlight individual differences in the use of transcription conventions by transcribers, for example, the way in which, episode titles and airdates are recorded. The work of semantic indexing presupposes the existence of an *episode template, i.e.* a textual standard to which the target text should conform. The process of semantic indexing is thus one of text modification that attempts to emulate and apply the notion of *episode template* systematically. Whether based on experience, or following explicitly stated guidelines, the enactment of this process requires both knowledge of the organization of texts and computational techniques. In the process of semantic indexing, preparing a transcript for such extraction is accomplished in main three steps: *Content cleaning*, *Semantic Annotation* and *Indexing*, each with various sub-steps, the main features of which are described below.

Content cleaning is the process of textual adjustment that we have outlined above. For the *House Corpus*, it involved turning *html* documents with embedded transcripts into corpus-ready transcripts in various steps, some of which are reproduced in Table 1. The process starts with the retrieval of the *source text* (*Point 1* in Table 1), which is achieved using Jsoup API (1), and subsequently proceeds with the cleaning process itself. The information contained within the <Title>tag of the HTML document is not standardized; each URL may store information differently. Table 1 (*Point 2*) shows five examples of different formats within the <Title> tag.

1. **Fetch the content of each URL.** Content is an HTML document.
2. **Extract episode title, season# and episode# by parsing the <Title> tag of HTML document.**
   a <title>House MD - 1.01 Pilot - House Transcripts**</title>**
   b <title>House MD – 4.13 No More Mr. Nice Guy - House Transcripts**</title>**
   c <title>House – S. EE TTTTTTTTTTTT - House Transcripts**</title>**
   d <title>MD - S.EE TTTTTTTTTTTT - House Transcripts**</title>**
   e <title>S. EE – TTTTTTTTTTTT - House Transcripts**</title>**
3. **Extract the main article from the HTML document of each URL**
   At this point, the HTML document contains transcript along with boilerplate text (advertisements, comments, template, navigational elements and other types of unrelated information).
4. **Extract the “Original Airdate” from the main article**
   Like title, the original airdate is also not standardized. The following are some examples of different formats of air dates from different URLs:
   a) Originally aired Apr 4 2006
   b) Originally Aired MMM DD YYYY
   c) Original Air Date on MM DD YYYY
   d) Original Air Date: : MMMM DD YYYY
   e) Original Air Date: MM DD YYYY
5. **Extract author(s) of the episode by standardizing the non-standardized string “Written by” to “Written by:” string**
6. **Remove unnecessary lines from the main article e.g. disclaimer messages**

Table 1
Steps in Content Cleaning.

Technically speaking, we can summarise the process involved as follows. First of all, dashes “–“ (HTML code &#8211) in the title are replaced with the minus “–“ (HTML code &#45) sign as some URLs contain dashes and some minus signs within the <Title> tag. Afterwards, if the <title> tag contains the strings “MD -” or “House -”, the title of the episode is reduced as a substring starting at index of ("-")+2 and ending at index of ("- House")-1. Otherwise, it is reduced as a substring starting at index of 0 and ending at index of ("- House")-1. At this point, the title string from some URLs’ content could contain dashes and dots (with spaces). If dashes are found they are removed from the string, whereas if dots with spaces are found they are replaced with dots. The title of the episode is extracted as a substring starting at index 4. The Episode # marker is extracted as a substring starting at index 0 and ending at index 1, while the season # marker is extracted as a substring starting at index 2 and ending at index 4.

The transcript is then extracted from the HTML document (*Point 3*) using Boilerplate API (Kohlschütter *et al*. 2010: 3), which provides algorithms to detect and remove the Boilerplate text/content around the main textual content of a web page. Other forms of standardization are then applied. For example, *Point 4* in Table 1 relates to the standardization of months as MMM (Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec) as compared with spellings, and above all misspellings, of months found in the main articles of URLs which included: *Janu, Febu, Marh, Apri, May, June, July, Augu, Sept,* *Octber, Nove, Dece, January, Feburary, March, April, May, June, July, August, September, Octobor, November, December*. Likewise the original airdate is standardized by replacing all cases as “Originally Aired:” and extracted as an index of (“Originally Aired:”)+2 while the date was changed from the MMDDYYYY format to the DDMMYYYY format.

The next stage in the *Semantic Indexing process* relates to *Semantic Annotation* using *Named Entity Recognition (NER)*. The latter is an information extraction task concerned with finding textual mentions of entities belonging to predefined categories, such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages and so on. NER systems take documents either in the form of blocks of plain text or, more directly, as URLs and transform them into annotated text. In fact, a modified version of *DBpedia Spotlight* was used. *DBpedia* (Lehmann *et al*. 2015) is designed, using the techniques associated with the *Semantic Web* (Berners-Lee *et al*. 2001), to extract structured content from the information created as part of the Wikipedia project. The structured information generated from Wikipedia pages is publicly available on the Web. *DBpedia* allows expert users to *semantically* query relationships and properties associated with Wikipedia resources, including links to other related datasets. As a NER, *DBpedia Spotlight* (Mendes *et al*. 2011) associates (i.e. links) Wikipedia resources to plain text.

Two aspects of the use of *DBpedia Spotlight* need to be highlighted. The first relates to *Transcript annotation*. Given the considerable time required for *DBpedia Spotlight* to annotate large documents, each transcript was split into multiple text blocks of about 20000 characters and then sent to *DBpedia Spotlight* for entity annotation. The resources thus obtained were subsequently merged. Once transformed into a *record* consisting of a transcript (or part of it), annotations, author, episode number etc. formatted in JSON format (Crockford 2006), each transcript was ready to be indexed. The second aspect relates to *Scene-wise annotation*, a defining feature in the *House Corpus Project*, which requires the possibility for each scene in an episode transcript to be extracted as a separate entity. The method used is a *Regular Expression* of the form *?((?i)cut .*?)\\]|CUT TO: which, translated, relates to any characters followed by the string “cut” or “Cut”, followed by more characters and a closing square bracket “]”, or just the string “CUT TO:”). Scene annotation is much slower than transcript annotation so that for larger corpora (not the case with the current corpus), the *DBpedia spotlight* service would need to be hosted on a local server for shorter delays.

Indexing is the final stage. This is a procedure whereby a Search Engine creates indices for records, thus allowing it to carry out searches more efficiently (https://en.wikipedia.org/wiki/Search_engine_indexing). For this, we used *Elasticsearch* (Gormley *et al*. 2015), a popular search engine. Developed in Java, *Elasticsearch* is released as open source under the terms of the Apache License. Based on *Lucene*, a free, open-source information retrieval software library, it is distributed, which means that indices can be divided into shards (*i.e.* partitions) and each shard can have zero or more duplicates (by default three for backup and other purposes). Thanks to these features, *Elasticsearch* provides near real-time search capabilities using an HTTP web interface which can be accessed by multiple users. After performing entity annotation, the JSON formatted documents were indexed into an *Elasticsearch* server hosted at the CNR Palermo, Italy (*http://openmws.itd.cnr.it*). A final consideration is the fact that indexing is such to allow the exclusion of some parts of the records from the indexing process. Thus, before indexing, it is essential to determine the right mapping for the index (JSON structure where the searchable fields, data types and sub types of fields are declared). For the *House M.D.* series, the default mapping of *Elasticsearch* was used whereby all the fields are set as analysed (*i.e.* searchable). However, *separate* indexes were created for full transcript documents (episodes) and split documents (*i.e.* those based on scenes).

3. Part 2: Scene Management and Annotation

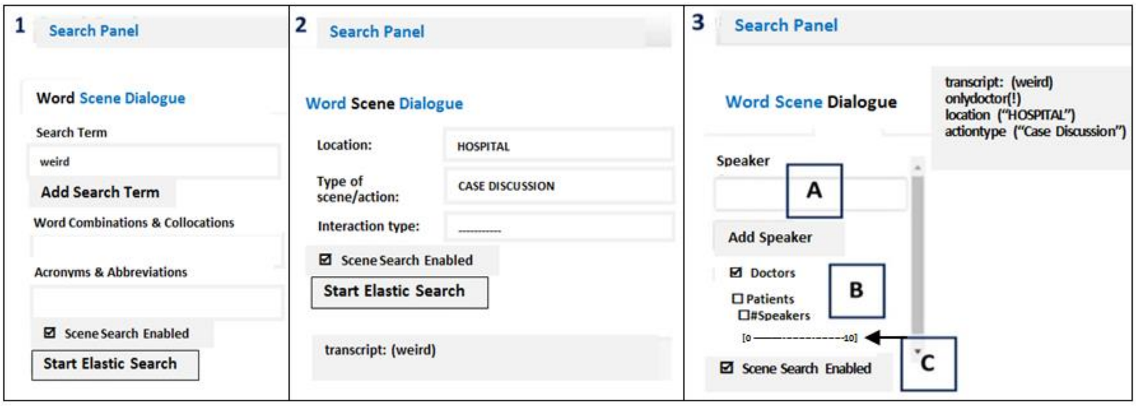
So far, the major focus in *House Corpus Project* has been on encouraging the capacity of university students, many in the very first years of degrees in language studies, to explore the grammar of English in ways that extend beyond the very basic frameworks acquired during years at school. This is achieved by encouraging engagement with the functions of specific lexicogrammatical structures in the scripted discourse of a well-known TV series. As well as supporting *Search functions*, the interface is also designed to allow students to perform further annotation of the corpus under the guidance of teachers. In a project designed to encourage participation in the manual annotation of corpora, the planning of scene-level indexing and of functionalities ideally needs to be built on the premise that the division of the 177 episodes into 6000-plus scenes, carried out in the preliminary stages of the project, opens up the possibility of creating maps of *scene types*. Intuitively, our experience of TV medical drama series suggests the following sequence of events: 1) a person is unexpectedly taken ill and rushed to hospital; 2) the patient is stabilised and the doctors attempt to establish the cause of the illness; 3) complications such as a condition’s rarity or concealment of information lead to improper diagnosis; 4) the true cause is eventually uncovered (in this TV series by Dr. House) and the case resolved; 5) the patient, from being on death’s doorstep, miraculously recovers and lives happily ever after.

The likelihood that different discourse structures will operate in different parts of an episode will be apparent, even from this basic sketch. For example, we would expect the present tense verb form *faints* to appear as part of the “stage directions” of an opening scene in which a character falls ill but for the past tense verb form *fainted* to appear in a history-taking and patient examination scene, shortly afterwards, where doctors get to grips with what actually happened to the patient. This pattern does in fact emerge: the form *faints* appears in *Scene 1 of Episode 9, Season 7*, and, as predicted, in a stage direction, while *fainted* appears early on in three episodes (*Season 3, Episode 18 Scene 03; Season 7, Episode 03, Scene 03; Season 8, Episode 14, Scene 07*). However, intuition is not enough to explain why *faints* also occurs in the resolution phase of an episode (*Season 2, Episode 16, Scene 25*) and *fainted* occurs as part of the complication phase (*Season 6, Episode 20, Scene 18*).

While word searches, as the *faint* example show, are a basic premise for the mapping of the various scenes, it is useful to turn matters around and make a scene search the starting point for discovering, for example, the list of verbs typically used in a specific *type* of scene, regarding which it is much harder to make intuitive predictions. Such maps are likely to be useful in supporting the work of various categories of potential users: apart from specialists in media discourse (Baldry 2016), they include all those interested in medical discourse, not just students and teachers of medical English, but also researchers and others developing or participating in specialist classes for medical translation and interpreting (Bianchi 2015). Furthermore, a typology of scenes appears to be an essential prerequisite for the efforts to construct a dialogue level of access, which, in its turn, is likely to be of benefit, for example, to those working in fields such as pragmatics and multimodality. However, in keeping with our primary goal of assisting student annotators in the discovery of discourse patterns within teacher-led projects, the focus has been on providing functionalities that make such manual annotations possible.

Put another way, the interface had to be as intuitive as possible, simplifying the *how-does-it-work* aspects of searching and annotating the corpus, while at the same time encouraging the desire to use the tool as a way to reflect on how the grammar of English is actually used in the production of discourse. To this end, though separate, the interface’s *Search* and *Annotation* functionalities are essentially specular, making it easy for students to test out the annotations they make immediately, all part of the process of encouraging discussion of their results with others, a vital aspect of the interface’s capacity to stimulate identification of distinctive discourse patterns.

For the purposes of illustrating the interface’s characteristics, we will first illustrate the *Search* interface, before describing the corresponding functionalities in the *Annotation* interface. As the first column in Figure 1 shows, the *Search Panel* interface allows selections to be made in terms of individual words or expressions made up of more than one word (*Word Panel*) that can be searched for in terms of the type of scene in which they appear. The second column in Figure 1 shows the searchable scene characteristics available (*Scene Panel*) relating to the way discourse is shaped and constrained by: a) *Location Type*, e.g. taking place within a hospital setting or elsewhere; b) *Event Type*, e.g. involving patient examination and history-taking, surgery, or, as shown in the example, a case discussion; c) *Interaction type* – currently restricted to scene closures (Coccetta 2019).



Search Term: weird
☑ Scene Search Enabled

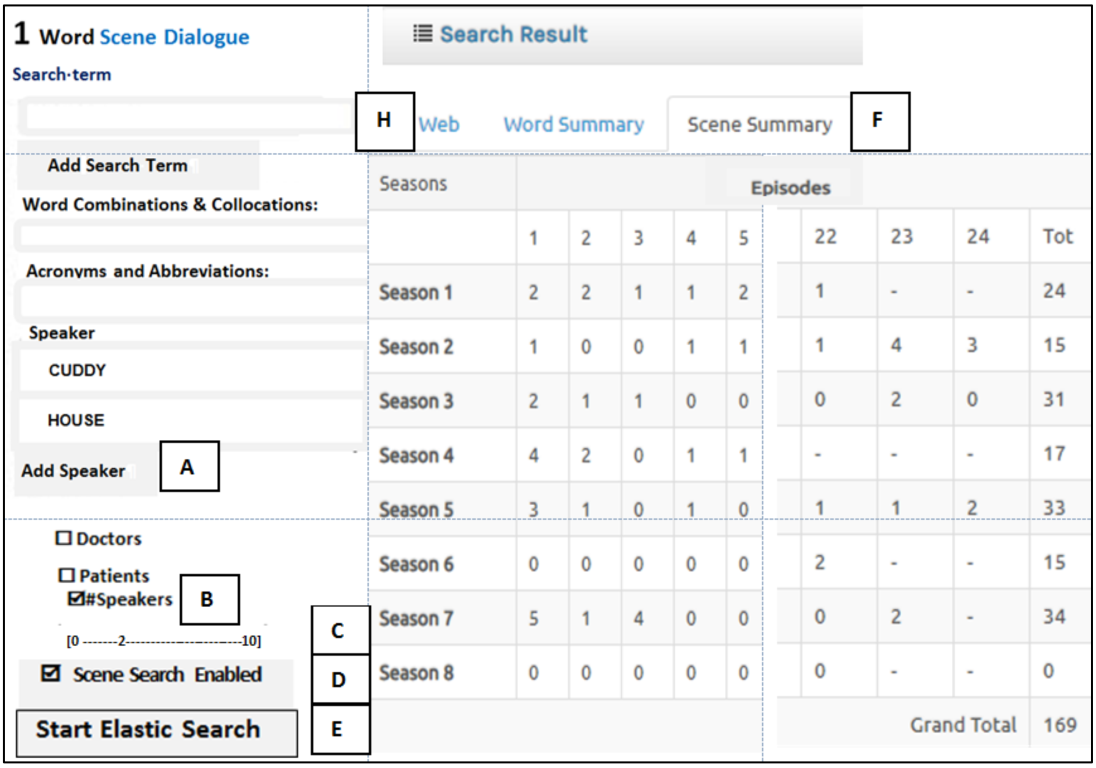
Location: HOSPITAL
Type of scene/action: CASE DISCUSSION
Interaction type: ---------
☑ Scene Search Enabled
transcript: (weird)

transcript: (weird)
onlydoctor(!)
location ("HOSPITAL")
actiontype ("Case Discussion")

☑ Doctors
□ Patients
□#Speakers
[0 ----------- 10]
☑ Scene Search Enabled

Figure 1
The three-part Search Panel.

The third column in Figure 1 shows a final panel (*Dialogue Panel*) relating to the interactants in the discourse. As the boxed letters show, this allows the user, for example, to select scenes in terms of: (a) specific speakers (*Box A: Speaker*); (b) categories of speakers (*Box B: Speaker Category*); (c) number of speakers in a scene (*Box C: Speaker Number*). As the first column in Figure 2 further illustrates, an entry for CUDDY and HOUSE in the *Speaker* textbox, requires the use of the *Add Speaker* function (*Box A*), plus selecting the *Speakers Box* (*Box B*), setting *0-10 Slider* to *2*, (*Box C*) and finally selecting the *Scene Search Enabled* box (*Box D*). This is all that is needed, apart from clicking the *Start Elastic Search* button (*Box E*), to identify the 169 scenes in which the *only* interactants are House and Cuddy. Cuddy is House's boss and there are many memorable scenes in which they confront each other alone so that, an expert user will want to learn more about the distribution of these scenes across the series. This function is carried out by the *Scene Summary* tool (*Box F*) illustrated in the second column of Figure 2. This generates a table which, although presented here in a clipped form for reasons of space, still identifies fluctuations in scene counts across seasons for this pair of characters and, indeed, shows that this type of scene disppears in the very last part of the series. To understand why, the user can check up on each individual scene using the *Web* tool (*Box H*).



Speaker
CUDDY
HOUSE

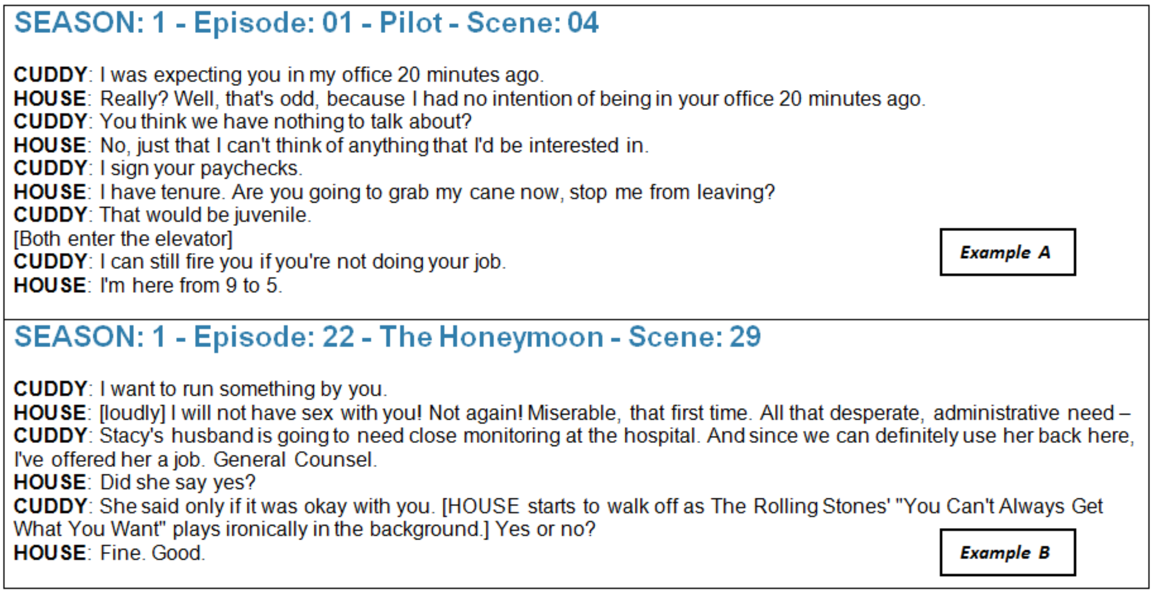
☐ Doctors
☐ Patients
☑ #Speakers
[0 -------2-----------------------10]
☑ Scene Search Enabled

Seasons	Episodes								
	1	2	3	4	5	22	23	24	Tot
Season 1	2	2	1	1	2	1	-	-	24
Season 2	1	0	0	1	1	1	4	3	15
Season 3	2	1	1	0	0	0	2	0	31
Season 4	4	2	0	1	1	-	-	-	17
Season 5	3	1	0	1	0	1	1	2	33
Season 6	0	0	0	0	0	2	-	-	15
Season 7	5	1	4	0	0	0	2	-	34
Season 8	0	0	0	0	0	0	-	-	0
								Grand Total	169

Figure 2
A Scene Search relating to two named interactants.

Additionally, thanks to the work on annotation undertaken by student annotators it is now possible to select scenes in terms of *Character Groups*, for example, scenes, which include doctor-only verbal interactions or scenes characterized by doctors' interactions with patients. As the *Scene Summary* functionality in the *Search Results Panel* in Figure 2 shows, the distribution of such scenes during the unfolding of the TV series varies considerably. The combination of this functionality, and the *Character Groups* functionality, sets up the possibility for teacher-led projects to be carried out that are sociolinguistic in nature and which might well be concerned with speaker distributions and the reasons for such variations in the various episodes and at different points in the overall TV series. While the tools already available are enough to enable such a project to be undertaken, other projects will require adjustments to the interface. For example, Figure 3 illustrates the fact, mentioned above, that each of the scenes identified in Figure 2 can be accessed via the *Web* functionality, the leftmost option in the *Search Result Panel's* main menu and marked as *Box H* in Figure 2. The scenes that Figure 3 reproduces are the first (*Example A*) and the last (*Example B*) of the Cuddy/House face-offs in the series' first season. Both examples in Figure 3 illustrate the constantly conflictual relationship existing between these two characters mentioned above that constitutes a major source of entertainment in the series as in other TV series (Baldry 2016). In this respect, *Speaker initiation* is high on the *to-do* list as regards functionality development as the search (Figure 2) which detected 169 scenes involving Cuddy/House interactions does not currently distinguish between those initiated by Cuddy and those initiated by House, a distinction that may well reveal differences in the incidence and circumstances of their confrontations.

The search subpanels in *House Corpus Search Panel* interface can be used separately or in combination. For instance, *Case Discussion Scenes*, can be subcategorized into those occurring *within* a specific *Character Group* (e.g. doctors only) and those occurring *between* a specific *Character Group* (e.g. doctors) and a specific individual (e.g. a patient or caregiver) named in the *Speaker Box*. Equally, the *Public/Private* distinction helps clarify why some House-Cuddy confrontations take place before intimidated patients but others occur more privately. The *Word Summary* functionality reports the distribution of searched-for words. As Figure 3 shows, searches need not be lemma-based but in many cases benefit from the inclusion of words. Had the word *job,* which appears in both scenes in Figure 3, been included in the search, the *Word Summary* tool would have shown the distribution across the series of the sixteen scenes with this combination of word and scene features. Additionally, in the individual scenes returned, the target word would have appeared in red as illustrated in many other examples in this article.



SEASON: 1 - Episode: 01 - Pilot - Scene: 04

CUDDY: I was expecting you in my office 20 minutes ago.
HOUSE: Really? Well, that's odd, because I had no intention of being in your office 20 minutes ago.
CUDDY: You think we have nothing to talk about?
HOUSE: No, just that I can't think of anything that I'd be interested in.
CUDDY: I sign your paychecks.
HOUSE: I have tenure. Are you going to grab my cane now, stop me from leaving?
CUDDY: That would be juvenile.
[Both enter the elevator]
CUDDY: I can still fire you if you're not doing your job.
HOUSE: I'm here from 9 to 5.

Example A

SEASON: 1 - Episode: 22 - The Honeymoon - Scene: 29

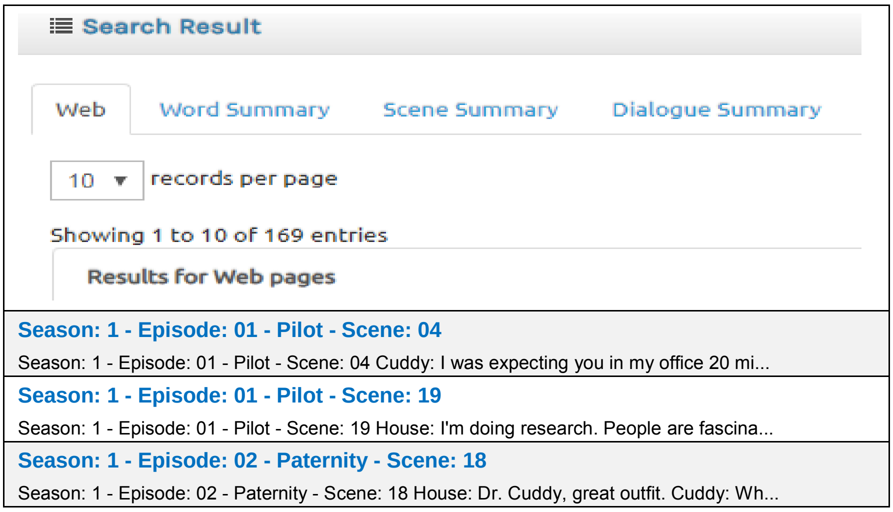
CUDDY: I want to run something by you.
HOUSE: [loudly] I will not have sex with you! Not again! Miserable, that first time. All that desperate, administrative need –
CUDDY: Stacy's husband is going to need close monitoring at the hospital. And since we can definitely use her back here, I've offered her a job. General Counsel.
HOUSE: Did she say yes?
CUDDY: She said only if it was okay with you. [HOUSE starts to walk off as The Rolling Stones' "You Can't Always Get What You Want" plays ironically in the background.] Yes or no?
HOUSE: Fine. Good.

Example B

Figure 3
Retrieved scenes: the first and last in this TV series where Cuddy and House clash.

As Figure 4 shows, access to specific scenes is made possible using the *Web* functionality, the first option in the *Search Result* Panel. This produces a list of scenes below the heading *Results for Web pages* ranked chronologically in the form of hyperlinks. Mouse selection of the final item in each hyperlink displays the scene in question. In this example, the words *Scene 03* in Figure 4, when clicked, will display the scene reproduced in *Example A* in the top part of Figure 3. *Dialogue Summary*, the final functionality in the *Search Result* Panel, is designed to quantify the frequency of specific types of exchange patterns but currently has the status of a *yet to be activated* option with characteristics to be defined on the basis of user feedback.

The division of the *Search Interface* into three panels is thus compatible with further customization and addition of new panels meeting the needs of teachers wishing to carry out specific student projects. Some of these have already been incorporated. Hence Figure 1 includes the possibility for searches to be carried out in relation to medical acronyms and abbreviations (Loiacono, Tursi, *this volume*). Equally, provision has been made for *Interaction types* to be included, currently implemented in terms of adjacency pairs (Coccetta 2019). Another project, involving dentistry students, is dealing with the annotation of behavioural verbs such as *cough* and *breathe* and will presumably lead to further adjustments of the interface.



10 records per page

Showing 1 to 10 of 169 entries

Results for Web pages

Season: 1 - Episode: 01 - Pilot - Scene: 04

Season: 1 - Episode: 01 - Pilot - Scene: 04 Cuddy: I was expecting you in my office 20 mi...

Season: 1 - Episode: 01 - Pilot - Scene: 19

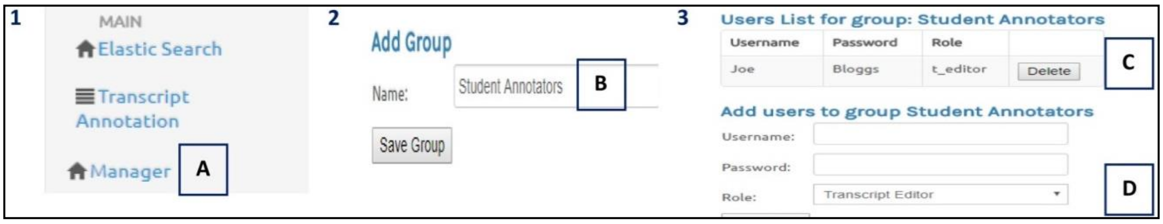
Season: 1 - Episode: 01 - Pilot - Scene: 19 House: I'm doing research. People are fascina...

Season: 1 - Episode: 02 - Paternity - Scene: 18

Season: 1 - Episode: 02 - Paternity - Scene: 18 House: Dr. Cuddy, great outfit. Cuddy: Wh...

Figure 4
List of scenes relating to Cuddy-House verbal exchanges.

Access to *Search* and *Annotation* functionalities is restricted through the *Profiling system.* The *Manager* functionality illustrated in Figure 5 shows the three steps required to provide groups of students with access to specific functionalities while excluding others. In the example shown, selection of the *Manager* functionality (first column, *Box A*), leads to a *Group Name* functionality (second column, *Box B*, in this case *Student Annotators*) followed by the addition (when so required) of a *Username* and *Password* (third column, *Box C*). Initially, this was a straight choice between *Searching* and *Searching and annotating* (*i.e. Transcript Editor*, third column, *Box D*), but a *Timepointing* functionality described below (see Figure 7a) was subsequently added. Further customisation, the result of user suggestions and analytics *i.e.* recordings of typical user-corpus interactions, will obviously be undertaken where appropriate.



Name: Student Annotators

Users List for group: Student Annotators

Username	Password	Role
Joe	Bloggs	t_editor

Add users to group Student Annotators

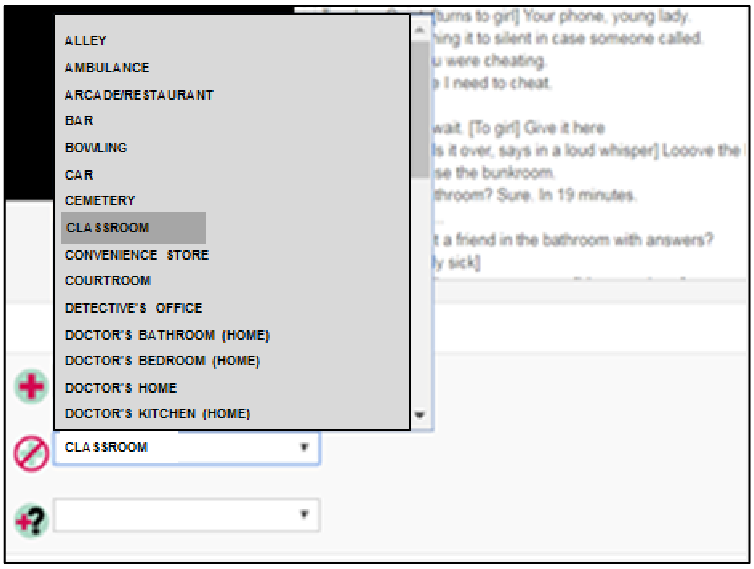
Username:

Password:

Role: Transcript Editor

Figure 5
Profiling system.

A partial illustration of the *Annotation Panel's* replication of the *Search Panel* interface is given in Figure 6, which exemplifies the icon-assisted possibilities for annotating specific scenes in relation to intra and extra hospital *Locations*, as well as undecided cases, i.e. those where a decision for annotators is hard to make. Having browsed through the scene in question (shown out of focus in the background), the annotator chooses from a list of over 50 extra-hospital settings used in this series, an easy choice in this case as the scene (*Scene 1, Episode 8, Season 1*) takes place in a classroom. The chosen option remains when the list is closed.



[turns to girl] Your phone, young lady.
hing it to silent in case someone called.
u were cheating.
I need to cheat.

wait. [To girl] Give it here
ls it over, says in a loud whisper] Looove the
se the bunkroom.
throom? Sure. In 19 minutes.

t a friend in the bathroom with answers?
y sick]

Figure 6
Some options for the annotation of scenes.

The access pathway to individual scenes is through a standard tree structure as illustrated in the various columns in Figure 7a. When the first column in Figure 5 is compared with the top-left hand corner of the first column in Figure 7a, it will be noted that the *Annotation interface* has changed. Thus, in this

CLASSROOM

configuration, in contrast to providing access to the *Transcript Annotation* functionalities illustrated in Figure 6, access is given to the very different *Timepoint Annotation* functionalities.

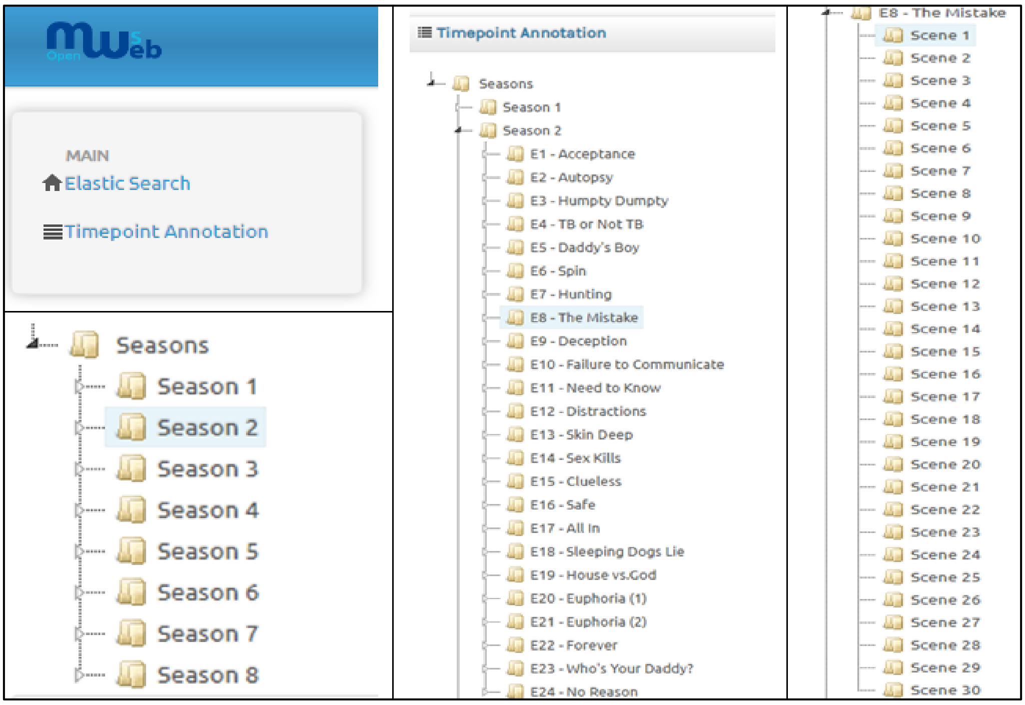
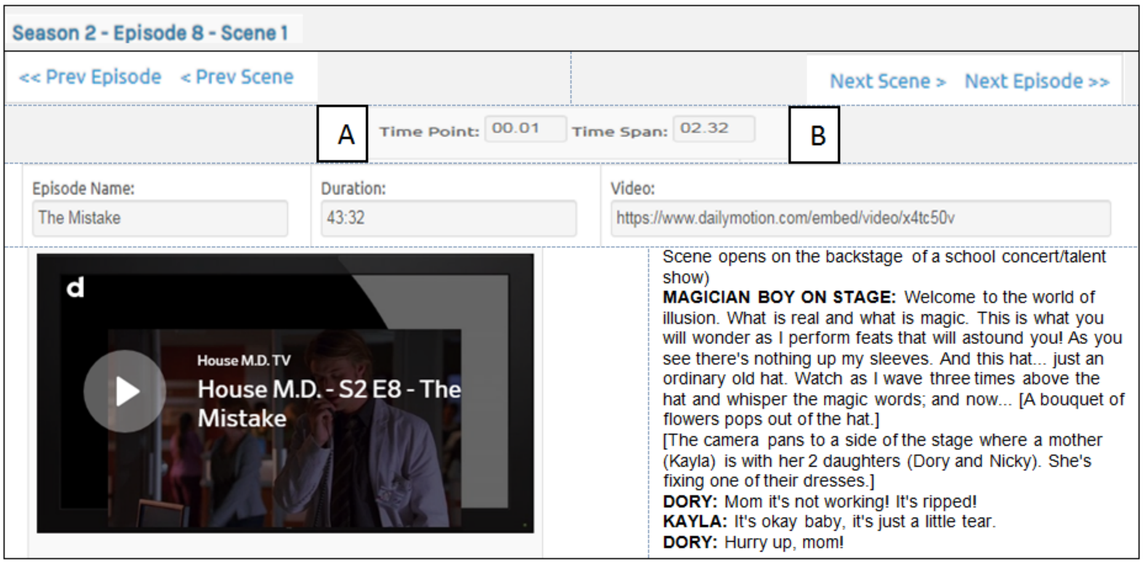


Figure 7a
Accessing annotation options for the link-up between scene reading and viewing.

The latter highlight the role student annotators can play in the work of associating scene transcripts with the corresponding video scene, as illustrated in Figure 7b. As the third column in Figure 7a shows, access to the scene shown in Figure 6 (*Scene 1 Episode 8, Season 2*) has been provided through the same access pathway but, as Figure 7b shows, the *Annotation* functionalities have changed. *Box A* in Figure 7b shows that the *Annotation* interface allows an annotator to indicate the point in the online video where a specific scene starts (in this case, the opening scene in the video), while *Box B* allows the scene's duration to be recorded.

In the initial state of research these annotations were limited to the *Annotation* interface. However, the now completed timepoint annotation work, undertaken entirely by students, was such to provide the data needed to support a corresponding *Search* functionality that allows an end-user to view, as well as read, the scenes that a particular search identifies. This takes the form of side-by-side comparisons of transcript and video versions of the same scene and, as Figure 7b shows, is achieved through links to the *DailyMotion* website (*https://www.dailymotion.com*). These links, which comply with the copyright restrictions stated on this website, encourage deeper investigation into the relationship between grammatical forms and their discourse functions, thanks to the possibilities of hearing as well as seeing the actors play out their lines.



Time Point: 00.01 Time Span: 02.32

Episode Name: The Mistake

Duration: 43:32

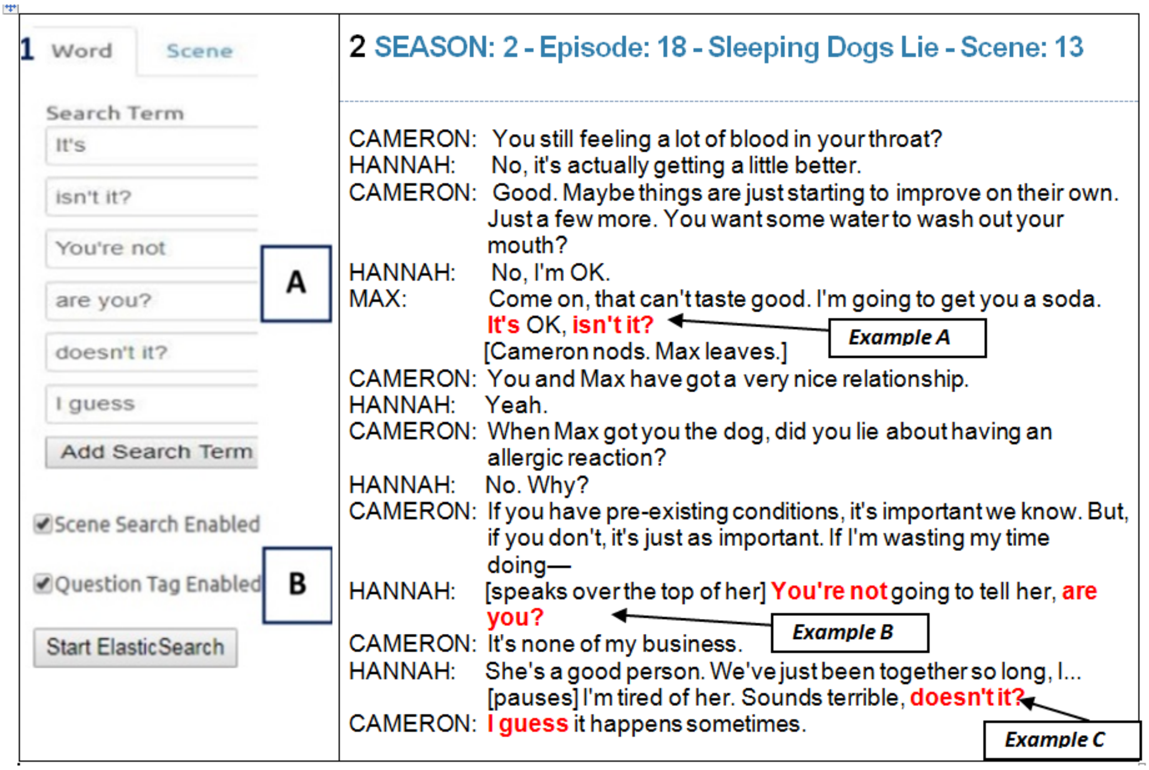
Video: https://www.dailymotion.com/embed/video/x4tc50v

Scene opens on the backstage of a school concert/talent show)
MAGICIAN BOY ON STAGE: Welcome to the world of illusion. What is real and what is magic. This is what you will wonder as I perform feats that will astound you! As you see there's nothing up my sleeves. And this hat... just an ordinary old hat. Watch as I wave three times above the hat and whisper the magic words; and now... [A bouquet of flowers pops out of the hat.]
[The camera pans to a side of the stage where a mother (Kayla) is with her 2 daughters (Dory and Nicky). She's fixing one of their dresses.]
DORY: Mom it's not working! It's ripped!
KAYLA: It's okay baby, it's just a little tear.
DORY: Hurry up, mom!

Figure 7b
Using the annotation options for the link-up between scene reading and viewing.

4. Part 3: Scene annotation and discourse analysis

TV drama series such as *House M.D.* offer many opportunities for a better understanding of the functions of scripted TV discourse in English. Given the need for high audience impact, the social and medical contexts chosen by screenwriters adopt a great variety of grammatical forms matched by an equally extensive variety of discourse functions. Take, for example, tag questions. Their use characterises oral discourse interactions in all varieties of English, although with surprising variations, in particular as regards frequency, in different parts of the English-speaking world (Tottie, Hoffmann 2006). While it is quite possible to find side-by-side spoken and written examples of tag questions in close-captioned *YouTube* films, detecting them may be likened to a hunt for microscopic needles in giant haystacks. The *House Corpus* instead finds examples easily and quickly. *Box A*, in the first column in Figure 8, shows how different tag and associated structures can be searched for, using, at the same time, the *Question Tag enabled function* indicated by *Box B*. This function eliminates tag-like forms which are not in clause-final position. As well as illustrating the system affordances, the examples in the second column of Figure 8 – multiple searched-for forms highlighted in red in a specific scene – also show various aspects of tag question patterns that can be used as a model by teachers in their illustration of the grammatical *vs*. discourse properties of tags in oral discourse in English. Thus, the first example shows a positive *anchor* (*it's*) and a negative *tag* (*isn't it*), while the second exemplifies the opposite polarity: a negative *anchor* (*you're not*) combined with a positive *tag* (*are you?*). The last example illustrates a negative tag (*doesn't it*) whose anchor is not another auxiliary but a lexical verb (*sounds*) with the typical subject ellipsis of spoken discourse. Additionally, the example highlights the discourse strategies involved in the use of tags; these often relate to seeking and providing reassurance.



Search Term: It's; isn't it?; You're not; are you?; doesn't it?; I guess

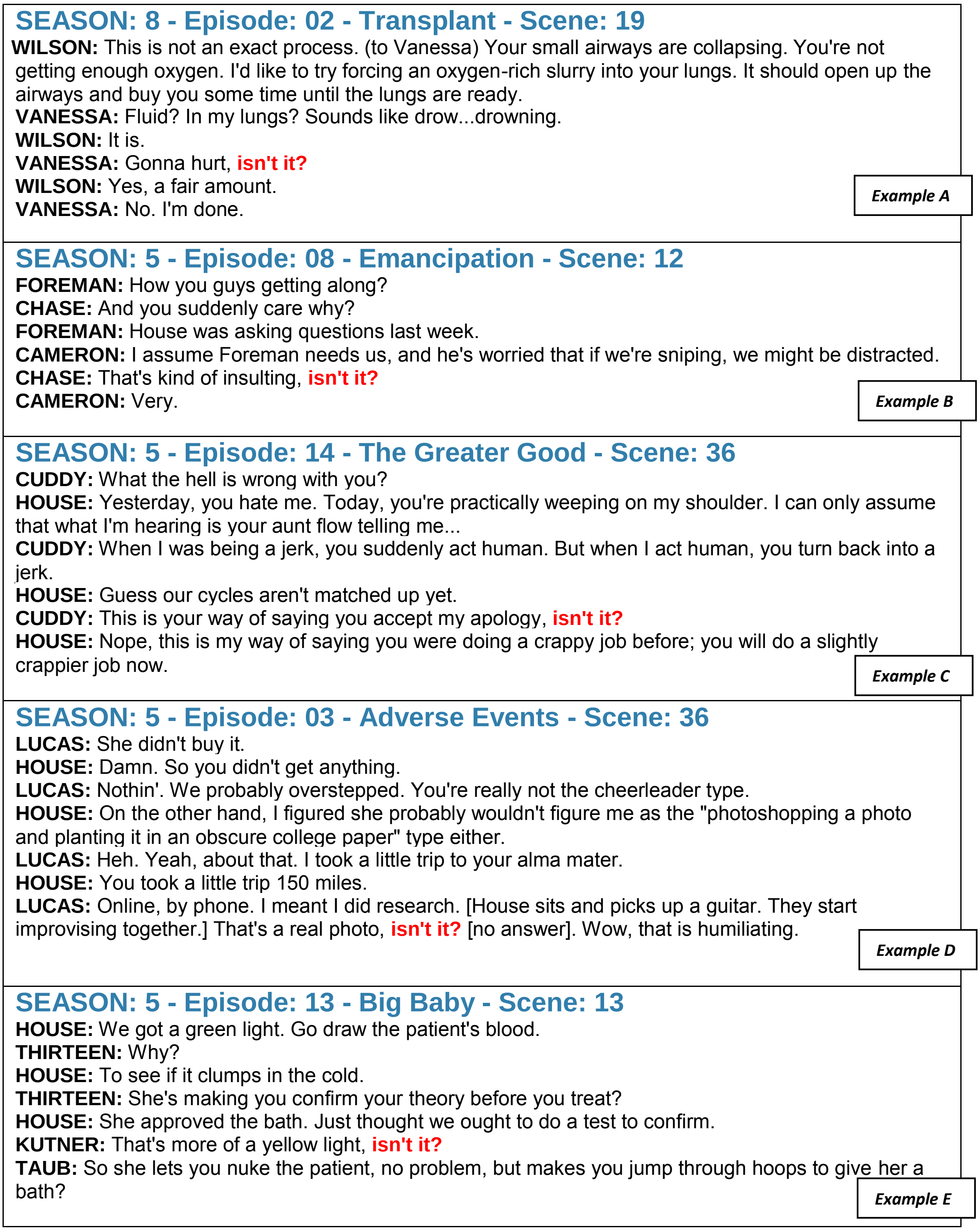
SEASON: 2 - Episode: 18 - Sleeping Dogs Lie - Scene: 13

CAMERON: You still feeling a lot of blood in your throat?
HANNAH: No, it's actually getting a little better.
CAMERON: Good. Maybe things are just starting to improve on their own. Just a few more. You want some water to wash out your mouth?
HANNAH: No, I'm OK.
MAX: Come on, that can't taste good. I'm going to get you a soda. It's OK, isn't it?
[Cameron nods. Max leaves.]
CAMERON: You and Max have got a very nice relationship.
HANNAH: Yeah.
CAMERON: When Max got you the dog, did you lie about having an allergic reaction?
HANNAH: No. Why?
CAMERON: If you have pre-existing conditions, it's important we know. But, if you don't, it's just as important. If I'm wasting my time doing—
HANNAH: [speaks over the top of her] You're not going to tell her, are you?
CAMERON: It's none of my business.
HANNAH: She's a good person. We've just been together so long, I... [pauses] I'm tired of her. Sounds terrible, doesn't it?
CAMERON: I guess it happens sometimes.

Figure 8
A multiple Question Tag search.

Indeed, most significantly, the added value that scene-based corpus searching brings lies precisely in the characterisation of the different *types* of reassurance sought and provided. In the scene shown in Figure 8, the first example is a request made by Max, the caregiver, for the doctor's agreement. In *Example A*, she seeks and obtains Cameron's reassurance (with a nod of the head) that no harm will be caused if the patient has a soft drink. While this reassurance relates to medical decision-making, the second type of reassurance in this scene regards non-disclosure of information, closely related to the issue which lies at the heart of this episode: professional integrity (*Example B*). Finally, the third example (*Example C*) relates to another type of reassurance concerned with a more personal and psychological plane, in which an experienced doctor allays the feelings of guilt and betrayal that a very sick patient, Hannah, has regarding the desire to leave Max, her companion/caregiver of many years.

Though all this is easily detectable thanks to scene-based searching, manual annotation can, of course, render discourse functions more easily detectable through specific annotation of question tag functions. However, the examples illustrated in the second column of Figure 8 show that, even without this higher level of annotation, scene-based corpus searches can go beyond typical corpus evidence relating to the frequency of specific lexicogrammatical forms and the ratio of negative to positive tags, as they provide easy access to the discourse functions that specific combinations of forms carry out in specific contexts. Indeed, as well as providing reassurance, tag questions also carry out other functions that demonstrate the need to *hear* and *see* their use in specific scenes in addition to examining them in transcript form. Thus, the pronunciation of a tag such as *"do you?"* – usually glossed as a stressed form when in utterance-final position – will in fact enact differing degrees of markedness according to the speaker's emotional state. Given the nature of drama in general, and House's relationships with his female boss in particular, we can expect that rebuttals rather than reassurance will prevail, as they are part of the conflicts that drive the drama in this TV series along. However, we can never be sure how this will be done. Thus in *Example C* in Figure 9, Cuddy, fishing for a compliment, meets with a rebuttal enacted by the colloquial form *Nope*. In other words, the 'grammatical' expectation, within the turn-taking system of oral and scripted discourse, for tag questions to cue dialogue partners to reply to the question with either a tag-based form of reassurance (e.g. *Yes it is*) or rebuttal (e.g. *No it isn't*) is not always fulfilled. Indeed, none of the take-ups in Figure 9 illustrate the *No, it isn't/ Yes, it is* pattern typically prescribed in rule-based 'grammar' lessons. *Example A* is the closest to such a pattern. It is perhaps easy to accept a response such as *Very* (Figure 9, *Example B*) as a legitimate and elegant breach of such rules, as this provides a strong form of reassurance. Nevertheless, it is the evasiveness of the final two examples that is particularly striking, so much so that, as *Example D* in Figure 9 shows, the original transcriber was so surprised that he or she wrote the bracketed words *[no answer]* immediately after the *isn't it* tag. Indeed, in contrast to the final example, Figure 9, *Example E* – where the *listener* takes evasive action and declines to respond to the tag question – *Example D* in Figure 9, is, instead, an instance of self-directed talk, a case where the current speaker breaks the next-speaker selection rule associated with tag questions by continuing to talk. Indeed, the speaker, shocked by the photo, is seeking self-reassurance, not reassurance from others. Within a manual approach to annotation, the functions of these four types of reassurance – that we may gloss as *medical, professional, psychological* and *self-referencing* – can be annotated with functional labels and subsequently searched for.



SEASON: 8 - Episode: 02 - Transplant - Scene: 19

WILSON: This is not an exact process. (to Vanessa) Your small airways are collapsing. You're not getting enough oxygen. I'd like to try forcing an oxygen-rich slurry into your lungs. It should open up the airways and buy you some time until the lungs are ready.
VANESSA: Fluid? In my lungs? Sounds like drow...drowning.
WILSON: It is.
VANESSA: Gonna hurt, **isn't it?**
WILSON: Yes, a fair amount.
VANESSA: No. I'm done.

Example A

SEASON: 5 - Episode: 08 - Emancipation - Scene: 12

FOREMAN: How you guys getting along?
CHASE: And you suddenly care why?
FOREMAN: House was asking questions last week.
CAMERON: I assume Foreman needs us, and he's worried that if we're sniping, we might be distracted.
CHASE: That's kind of insulting, **isn't it?**
CAMERON: Very.

Example B

SEASON: 5 - Episode: 14 - The Greater Good - Scene: 36

CUDDY: What the hell is wrong with you?
HOUSE: Yesterday, you hate me. Today, you're practically weeping on my shoulder. I can only assume that what I'm hearing is your aunt flow telling me...
CUDDY: When I was being a jerk, you suddenly act human. But when I act human, you turn back into a jerk.
HOUSE: Guess our cycles aren't matched up yet.
CUDDY: This is your way of saying you accept my apology, **isn't it?**
HOUSE: Nope, this is my way of saying you were doing a crappy job before; you will do a slightly crappier job now.

Example C

SEASON: 5 - Episode: 03 - Adverse Events - Scene: 36

LUCAS: She didn't buy it.
HOUSE: Damn. So you didn't get anything.
LUCAS: Nothin'. We probably overstepped. You're really not the cheerleader type.
HOUSE: On the other hand, I figured she probably wouldn't figure me as the "photoshopping a photo and planting it in an obscure college paper" type either.
LUCAS: Heh. Yeah, about that. I took a little trip to your alma mater.
HOUSE: You took a little trip 150 miles.
LUCAS: Online, by phone. I meant I did research. [House sits and picks up a guitar. They start improvising together.] That's a real photo, **isn't it?** [no answer]. Wow, that is humiliating.

Example D

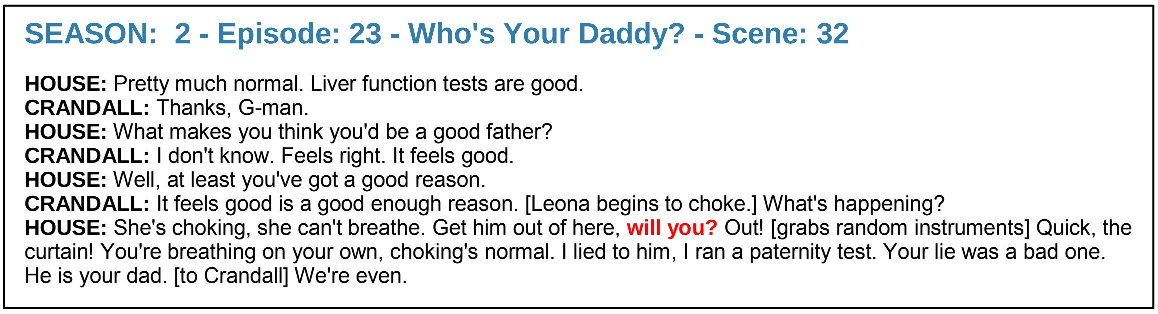
SEASON: 5 - Episode: 13 - Big Baby - Scene: 13

HOUSE: We got a green light. Go draw the patient's blood.
THIRTEEN: Why?
HOUSE: To see if it clumps in the cold.
THIRTEEN: She's making you confirm your theory before you treat?
HOUSE: She approved the bath. Just thought we ought to do a test to confirm.
KUTNER: That's more of a yellow light, **isn't it?**
TAUB: So she lets you nuke the patient, no problem, but makes you jump through hoops to give her a bath?

Example E

Figure 9
Contextualizations of the *isn't it?* tag question.

It could be argued that an interface specifically designed to look for *anchor* and *tag* sequences would represent an improvement over the current *Tag Question Search* function which merely allows searches for *tag questions* (and not their anchors) to be made. In this respect, a further consideration is that structures exist in English that have the same form and final position in utterances as tags. However, as the *JENNIFER: Stop it, will you?* example (*Season 7, Episode 20, Scene 22*) shows, such forms have no anchor. They are not a *You won't stop it, will you?* type of structure and do not express reassurance-seeking functions. On the contrary, they are typically demands for something to be done in moments of crisis or conflict and with a degree of insistence bordering on anger. If we add House as speaker into the search using the *Dialogue Panel* in the manner illustrated above in *Part 2*, it immediately becomes clear that four of the five examples of this type in the *House Corpus* are uttered by House and that this structure is associated with his role as team leader in medical emergencies, as Figure 10 illustrates.



SEASON: 2 - Episode: 23 - Who's Your Daddy? - Scene: 32

HOUSE: Pretty much normal. Liver function tests are good.
CRANDALL: Thanks, G-man.
HOUSE: What makes you think you'd be a good father?
CRANDALL: I don't know. Feels right. It feels good.
HOUSE: Well, at least you've got a good reason.
CRANDALL: It feels good is a good enough reason. [Leona begins to choke.] What's happening?
HOUSE: She's choking, she can't breathe. Get him out of here, **will you?** Out! [grabs random instruments] Quick, the curtain! You're breathing on your own, choking's normal. I lied to him, I ran a paternity test. Your lie was a bad one. He is your dad. [to Crandall] We're even.

Figure 10
Contextualization of will you?

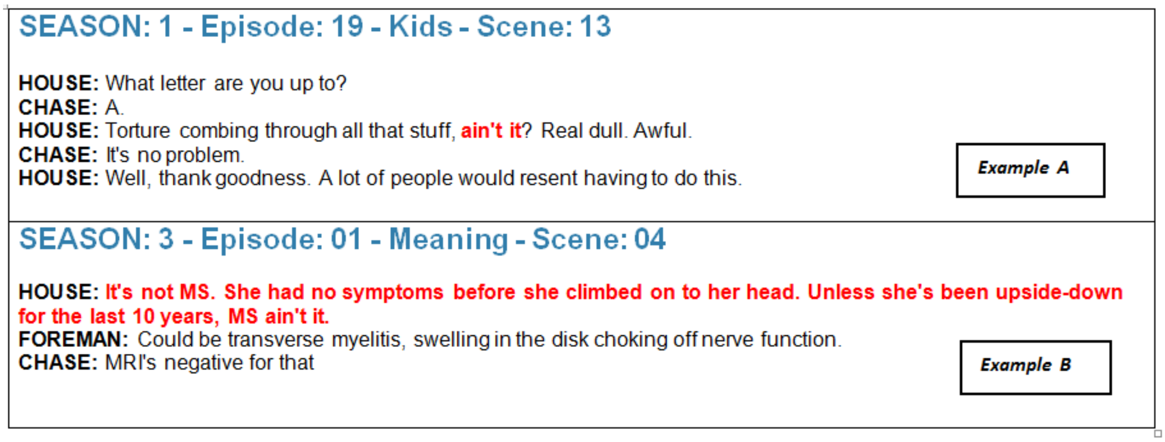
However, there is considerable complexity associated with detecting *anchors*, and highlighting them for easy user identification. Tags are constructed from a closed set of grammatical items, listed in Table 2, consisting of: (a) auxiliary and modal verbs with either negative or positive polarity (a distinction marked in Table 2 with a slash) and (b) personal pronouns plus *there* and *one*.

Am/Ain't	Can/can't	Did/Didn't	Is/isn't	Was/wasn't
Are/Ain't	Could/couldn't	Had/hadn't	Must/mustn't	Were/weren't
Is/Ain't	Do/don't	Has/hasn't	Shall/shan't	Will/won't
Are *or* Am/aren't	Does/doesn't	Have/haven't	Should/shouldn't	Would/wouldn't

Table 2
Tag Question Set.

However, their anchors belong to a far less restricted set of grammatical structures (see *Example C* in Figure 8). Indeed the anchors for *do, does* and *did* tags, and their negative counterparts, belong to an open-ended class of lexical items. Moreover, in some cases, no anchor will be present as a result of ellipsis (see *Example A* in Figures 9 and 11). The last line in the first column of Table 2 also includes the tag *am I* as in *I'm not here, am I*. Like the *ain't* form, this breaks with the basic pattern as the 'reverse' form, *I'm here, aren't I*?, requires different morphological selections compared with other cases where the order of negative and positive forms can, in theory, be swapped freely. Whether they *are* is another matter: some forms such as *can't* are so frequent that they appear in every episode of *House M.D.*, while the forms *mustn't* and *shan't* appear in none, thus *de facto* reducing the number of potential tag question *type:token* ratios to be tabulated and possibly presented, for example, in classroom teaching.

It will always be possible to find ways of automatically detecting and highlighting the ties between anchors and tags, and thus provide a resource that illustrates significant patterns of cohesion in oral discourse. However, as further suggested below in the *Discussion Section*, within the logic of student engagement with annotation advocated in this paper, it seems more appropriate to carry out manual annotation of anchors that encourages students to explore the 'conflict' between 'grammar' rules and 'discourse' rules and understand that they are two interdependent aspects of the overall process of meaning making.



SEASON: 1 - Episode: 19 - Kids - Scene: 13

HOUSE: What letter are you up to?
CHASE: A.
HOUSE: Torture combing through all that stuff, **ain't it**? Real dull. Awful.
CHASE: It's no problem.
HOUSE: Well, thank goodness. A lot of people would resent having to do this.

Example A

SEASON: 3 - Episode: 01 - Meaning - Scene: 04

HOUSE: **It's not MS. She had no symptoms before she climbed on to her head. Unless she's been upside-down for the last 10 years, MS ain't it.**
FOREMAN: Could be transverse myelitis, swelling in the disk choking off nerve function.
CHASE: MRI's negative for that

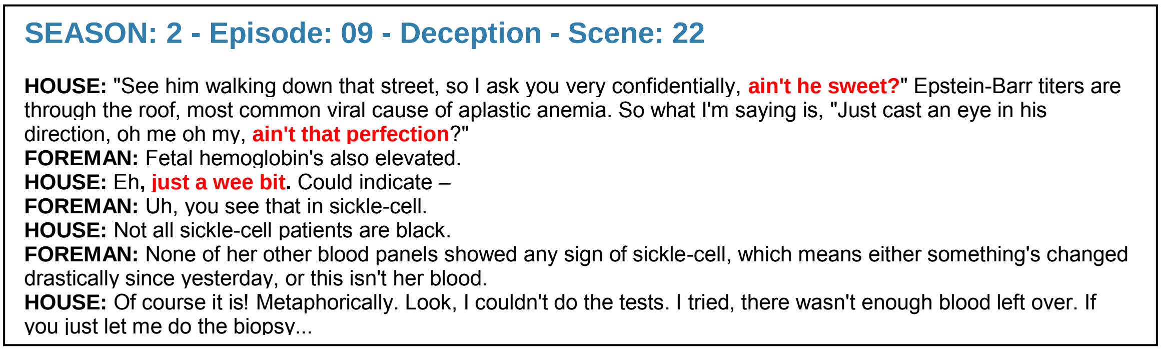
Example B

Figure 11
Contextualization of *ain't it.*

Given the limited resources so far available in this project, of more immediate concern have been the investments required to link up transcript scenes with their corresponding video scenes. Even so, it is worthwhile re-affirming the significance of prosodic features in distinguishing tag *look-alikes* from the real thing and hence the fundamental importance of comparative side-by-side readings and viewings that specialised multimedia corpora like the *House Corpus* make available. Alongside forms as such as *isn't it?*, considered 'standard' forms in oral discourse across many varieties of English, there are other forms viewed as substandard whose credentials are rarely presented in English language lessons in schools. As Cheshire (1991) points out, *ain't* is a frequent non-standard form of American and British English, not inflected for person and number, with five 'standard English' equivalents: *haven't, hasn't, (a)m not, aren't* and *isn't*. Figure 11 presents two examples of *ain't it* in the *House Corpus*, the first of which (*Example A*) is a tag question while the second

(*Example B*) is not. Viewings of the two scenes illustrated in Figure 11 show completely different intonation and stress contours that are in keeping with the different functions performed.

Figure 12 shows a scene where *ain't,* eschewed in written discourse in English, is once more used, this time with reference to a jazz era song: *Ain't he sweet.* Like its stablemate, *Ain't she sweet,* it epitomises the freedom of expression and defiance vis-à-vis expected grammatical and discourse strategies that characterise all songs. The song has been sung in many parts of the English-speaking world and recorded by a multitude of singers, including such household names as Nat King Cole, Frank Sinatra and the Beatles, promoting *ain't* as a form characteristic of informal varieties of English. It was thus only to be expected that Milton Ager and Jack Yellen's lyrics (https://lyricsplayground.com/alpha/songs/a/aintshesweet.html) would come to be woven into the *House M.D.* series. Figure 12 reproduces the scene where the devious and deviant Dr. House sings two lines from this song mixing medical lexis with jazz-era colloquialisms, thereby breaking the conventions of case discussions and differential diagnosis – as well as illustrating the need for corpus studies to find ways of detecting intertextual references. Naturally, manual annotation is one such way.



SEASON: 2 - Episode: 09 - Deception - Scene: 22

HOUSE: "See him walking down that street, so I ask you very confidentially, **ain't he sweet?**" Epstein-Barr titers are through the roof, most common viral cause of aplastic anemia. So what I'm saying is, "Just cast an eye in his direction, oh me oh my, **ain't that perfection**?"
FOREMAN: Fetal hemoglobin's also elevated.
HOUSE: Eh, **just a wee bit.** Could indicate –
FOREMAN: Uh, you see that in sickle-cell.
HOUSE: Not all sickle-cell patients are black.
FOREMAN: None of her other blood panels showed any sign of sickle-cell, which means either something's changed drastically since yesterday, or this isn't her blood.
HOUSE: Of course it is! Metaphorically. Look, I couldn't do the tests. I tried, there wasn't enough blood left over. If you just let me do the biopsy...

Figure 12
Contextualization of *ain't he sweet* and *ain't that perfection.*

Songs and singing are essential to any TV drama series. *House M.D.* is no exception. *House M.D.*, like many TV series, is characterised by the constant presence of music and song, in its affirmation of American language and culture (Law 2015). As it grows, the *House Corpus* will assist understanding of how grammatical and interactional selections are underpinned by awareness of, and references to, shared culture, songs being just the tip of this iceberg. Quite apart from the possibilities of detecting scenes that include songs, there is a need to reflect on the *textual* functions of songs, and more generally voice prosodics, within TV dramas, a matter that will be investigated in a subsequent phase of research in the *House Corpus Project.* In the *House M.D.* series, linguistic and cultural aspects are constantly referenced and celebrated as is further underscored in the scene reproduced in Figure 12 with its use of the expression *a wee bit* – universally associated with Scottish speakers – all evidence of the fact that, if all aspects of discourse are to be represented, corpus studies need to entertain the bigger picture of what is culturally shared in the English-speaking world, a picture for which word-based corpus searches are not noted.

5. Discussion

While the number of words spoken in the *House M.D.* has long been established at just under a million (Law 2015), the number of scenes is never mentioned – despite their centrality in any discussion of a TV series. Many type/token ratio analyses for words (Sinclair 1991; Butler 1997), obtained by dividing the number of different words (types) by the total number of words (tokens), have been produced. The procedure has been extensively critiqued with evaluations of a general nature such as Flowerdew's (2012, pp. 13-16) description of the difficulties of identifying types, as well as more specific assessments of their comparative potential in general *vs.* specialised corpus studies such as the work of McEnery *et al*. (2002) in relation to comparison of the BNC and the 100 Corpus of phone transcripts. A search for studies and critiques of type/token ratios for scenes in which the number of different types of scenes is divided by the total number of scenes in TV dramas will, on the other hand, simply draw a blank. Such ratios are the basis for the scene maps described above, a matter which raises the question as to what applications scene type/token relationships are designed to stimulate. There are many potential answers to this question, some involving purely didactic activities such as identifying scenes containing medical acronyms and thus clearly related to the lexical aspects of specialised L2 learning (Loiacono, Tursi, this volume); others instead might be concerned with research activities with no connection whatsoever to language learning or discourse analysis activities, for example, comparisons across different TV medical dramas of specific scene types such as those portraying medical emergencies which might be useful for TV critics. Obviously, there are strong affinities between language learning and discourse analysis activities. For example, corpus annotation of the type envisaged in the *House Corpus Project* obviously promotes active engagement with oral and written discourse in English in ways that encourage indirect forms of language acquisition (Krashen 1982). Many studies have, of course, suggested the significance of video in improving listening comprehension skills in a variety of teaching (Elk 2014), self-learning (Balcikanli 2010; Richards 2015; Takaesu 2017) and testing contexts (Lesnov 2017; Wagner 2010) as well as other more specialized contexts such as those concerned with the need for specific teacher training (Park, Cha 2013) or general reflection on the use of video in relation to the acquisition of listening and other comprehension skills (Bianchi 2015; Watkins, Wilkins 2011). Even so, to date, few research projects have contemplated the use of a corpus-based methodology that allows specific oral discourse features to be selected and practised with the advantages of precision and selectivity that corpus-based techniques bring. Some of these (Ackerley, Coccetta 2007, p. 353; Coccetta 2011) include multimedia corpus projects that address the cultural and social issues that we have mentioned above.

However, language learning is not what this project is about. Our concern is instead with defining scenes in ways that make them compatible with encouraging student engagement with CDA (critical discourse analysis) within the framework of corpus linguistics. This is the foundation stone on which the *House Corpus Project* is built and why the authors are concerned with the concept of functionality planning and investments in functionalities that bring about new forms of the empowerment that enhance such engagement.

How has such planning affected *House Corpus* R&D? Within the framework of functionality cost-benefit planning, genre selection was the first factor to be considered. The digital age has brought with it new affordances for the simultaneous side-by-side presentation of more substantial units of written and spoken discourse. For example, *Ted Talks* reinterprets the relationship between spoken and written forms in a way that goes beyond traditional subtitling as it allows users to display videos and their transcripts in the same window thus enabling viewers to watch a video and read its transcript simultaneously. Even so, the *Ted Talks* solution only offers: “monologic talk. The camera moves between long or close shots on the speaker, close shots on the projected slides, and long shots on the listening audience” (Bianchi, Marenzi 2016, p. 27). Given that variety is the spice of life, many users, students and teachers alike, will yearn to go beyond the *Ted Talks* ‘talk’ genre. Although as with many types of lecture, these talks are highly interactive, they do not illustrate the discourse features associated with interactional exchanges in English that characterise many oral discourse genres of English, exposure to which students enrolled in degree courses dealing with English language studies are in desperate need.

Scene analysis is a second example of functionality planning in which cost-benefit analysis was crucial. Our original division into scenes, as recorded in Part 1 of this paper, is based on references to scene cuts described in online transcripts (see also Law 2015) which thus provided a low-cost entry point for the project. However, defining where a scene starts and where it ends affects the way scenes are defined and quantified. Research promoting automatic scene detection has long recognised the difficulties of detecting scene boundaries (Ewerth, Freisleben 2004). Perceptions of what a scene is differ, a factor, which for better or for worse, constantly needs to be taken into account and, above all, explored in investigations of discourse in English. This explains our cautious use of the expression ‘6000-plus scenes’ when referring to the partial annotation of scenes subsequently carried out by students in the University of Salento in terms of typological features: *location type* (e.g. patient’s hospital room; medical lab); *event type* (e.g. differential diagnosis; precipitating medical event; patient examination) and *Character Group type* (e.g. doctor/doctor; doctor/patient; doctor/caregiver; patient/caregiver etc.). Indeed, the number of scenes has already increased thanks to manual annotation carried out by student annotators who have suggested splitting up scenes into smaller ones on the basis of the systematic application of these typological features. In whatever way a scene is defined, there will always be exceptions. For example, putting forward the idea that a scene is defined in terms of a change in location simply raises the question as to what is meant by a change in location and whether, for example, the frequent scenes in *House M.D.* which include multiple flashbacks are to be defined in terms of the *current* or *predominant* location. As such, from a methodological standpoint, promoting the scene to the status of a searchable but manually taggable unit is a liberating factor. At the very least, it enables students to modify the search results produced by allowing them to introduce *their* annotations about scene characteristics in compliance with the objective of promoting corpora as a way into CDA for undergraduate students.

A third example of functionality planning relates to compatibility with the *short course* and *in-spare-time* solutions. Thus, although corpus construction in general remains within the realm of advanced research, a few studies have described and discussed experiments that involve the participation of students. In one such project:

> participants were given access to specialized corpora of academic writing and speaking, instructed in the tools of the trade (web- and PC-based concordancers) and gradually inducted into the skills needed to best exploit the data and the tools for directed learning as well as self-learning. After the induction period, participants began to compile two additional written corpora: one of their own writing (term papers, dissertation drafts, unedited journal drafts) and one of "expert" writing, culled from electronic versions of published papers in their own field or subfield. Students were thus able to make comparisons between their own writing and those of more established writers in their field (Lee, Swales 2006, p. 56).

Such experiments typically rely on a substantial initial training period and are thus often directed to postgraduate students. This is incompatible with the realities of undergraduate training where CDA and corpus annotation cannot afford to overshadow other objectives. Within the framework of the further annotation of a pre-existing corpus, the *House Corpus Project* pursues a policy of creating micro-projects, that are easily manageable within a *to-be-completed-by-the-end-of-term* timescale, or where appropriate, even shorter periods. The major characteristics of this policy are:

1) Minimum-initial procedural training: learning how the system works requires at most a single live demonstration or a manual consisting of a few pages;
2) Targeting of very specific grammatical and discourse features;
3) Promotion of Teamwork: the model is designed for "group project work" among students in the early stages of their academic career; it enhances confidence through awareness that the annotations made add to the value of the corpus;
4) Customisation: the possibility of adding new annotational features that can subsequently be re-used by different groups for different tasks with minimal need for 're-tooling';
5) Teacher management: the teacher conducting a project has considerable control over the project thanks to *profiling tools* and *data analytics* that allow a teacher to monitor the progress of a group of students as well as each student individually.

The *House Corpus Project* envisages the addition of functionalities on an *as the need arises basis*. Indeed, the project depends on two inter-related aspects of interface management, namely the possibility of increasing the number of functionalities but also the adjustments that can be made to existing ones, which includes delegation of *decision-making* about such adjustments to teachers and/or students.

Clearly, this paper reports on the early stages of this project in which frequency data are not be available. Our curiosity is such, of course, that we, too, are eager to learn the ratio of intra- to extra- hospital scenes and whether scenes that occur in an extra-hospital environment are typically shorter or longer than scenes in an intra-hospital environment just as we would like to know the average length of a scene in this TV medical drama genre. Such knowledge would allow us to identify patterns and provide a basis for explaining why such patterns, and exceptions to them, occur. However, from the standpoint of functionality planning our interest lies elsewhere. In the initial stages of the project, as might be expected, the level of delegation was highly restricted. As the use of the *House Corpus* increases, so the pressure to delegate responsibility for the creation and management of functionalities also increases. Let us review these pressures in terms of functionality planning and what delegation of responsibility entails with some concrete examples.

If we return to the issue of speaker distributions within a university CDA short course project with a sociolinguistic orientation, we may note that it is already clearly possible, with the tools already existing, to carry out searches relating to the distribution of scenes *per episode*, *per season* and *per series* in the following ways:

1. *Numerically*: *i.e.* scenes with no speakers, a single speaker, two speakers and so on;
2. *Per individual: i.e.* scenes with specific characters named either in the metadata (e.g. speaker cues) or referenced in the discourse;
3. *Combinations* of these two parameters.

Note, however, that although the corpus is indexed in terms of individually named speakers, the current interface does not fully allow scenes to be identified in terms of speaker characteristics other than speaker name. Minor interface adjustments building on the student annotation functionalities already provided will make it possible to explore the power relationships implicit in interaction in terms of:

1. *Gender: e.g.* a project designed to annotate and explore the ratio of male-only scenes to female-only scenes;
2. *Professional and social standing*: e.g. a project looking into the construction of a *Category group* such as *caregivers* and the interactional expectations and realities associated with this category.

Thus, in the next stage of development, the intention is to create functionalities that allow a greater degree of delegation for a) teachers with respect to the system designers and b) students with respect to teachers in the construction of search categories. Thus, with a view to enabling *Gender* and *Cross category group* annotations, it is intended to:

1. provide the *Dialogue interface* with a *Speaker Group* function that allows new groupings of speaker names to be constructed;
2. allow a teacher to decide whether or not to make the *Speaker Group* function available to students in a project;
3. request students to determine the members of the *Speaker Group* in accordance with a specific project's objectives.

A similar pattern of delegation will likewise allow new annotational subcategories to be added to the pre-existing *Location type* and *Event type* parameters. While such changes require some rewriting of the interface rules, they are well within the bounds of possibility. On the contrary, a similar arrangement, creating a *Word Group* functionality, whereby users define and search for sets of related lexical items within the *Word* interface, would be a time-consuming IT task involving complex search rules and is thus currently not an option being taken into consideration. The issue of tag questions is, indeed, instructive as regards the cost-benefit ratio of investing in certain functionalities and not others in terms of the degree of delegation that can be achieved. Tottie and Hoffman (2006, p. 296) state that, when searching for "entire tags consisting of auxiliary, pronoun, and optional *n't,* we found a total of 200 different combinations, most of them occurring in very low proportions". Thus, as Part 3 has shown, from the standpoint of investment in learning experiences, delegating the solution to student annotators has many merits. The *Question Tag enabled* functionality currently only available in the *Search Interface* could be added to the *Annotation Interface,* based on a pre-established table of options, such as the one shown in Table 2. This would then allow *manual* annotation of *anchors* to be performed on a scene-by-scene basis by students as an end-of-term class project, with items from the list in Table 2 assigned to different groups. Of course, this raises the issue of the benefits that such a project would bring to the students in terms of exercising their CDA and corpus search & annotation skills, a matter that would have to be decided by the teachers overseeing such a project.

In the current stage of research, it is not entirely possible to predict which functionalities will be required, nor the benefits that the student engagement approach will bring as more data is required, in particular, as regards the value that has been added by associating scenes extracted from the corpus with the corresponding video scenes. The expectation, however, is that the answer to issues of functionality planning lies with data analytics as the recordings of user searches and annotations will provide a better guide to management aspects relating to the delegation, addition and modification of existing functionalities and the cost-effectiveness and benefits to students of further investment in new functionalities.

6. Conclusion

As the article reports, though indexed in ways described in Part 1, the *House Corpus* leaves open the possibility for annotations of a manual nature to be made to specific scenes in the TV series. Through a system of restrictive passwords and other controls, the interface is designed to allow university teachers to carry out specific annotation projects with selected groups of University students in which the scripted discourse of an entire TV series is explored with a view to adding annotations that enrich the value of the overall corpus. As such, while encouraging learning that relates to specific aspects of discourse in English, as illustrated in Part 3 with regard to the use of tag questions, the research reported, in keeping with the training and educational goals promoted by the institutions to which the authors are affiliated, is concerned with the development of online tools that exercise students' ability to acquire critical skills in the description of the discourse of written and spoken varieties of English through a hands-on approach to annotation. From the results so far obtained, promoting students' CDA skills through greater awareness of the characteristics and functions of corpora appears to be a viable proposition.

The project thus raises a basic question about the role of specialised corpora. Are they an end-product to be construed on a par with a printed dictionary for the purposes of consultation or are they to be seen instead as part of a collaborative learning experience in which the corpus itself is subject to the process of modification? From the exposition given above it is clear that the *House Corpus Project* is attempting to provide a strong stimulus in support of the view that specialised corpora can and should drive learning processes through student engagement with annotation and searching. Indeed the tag question example shows that the affordances created by hybrid forms of manual and automated annotation give a new twist to the term *blending learning*. From a procedural standpoint, the tag is identified and highlighted on the basis of abstract search rules enacted by a search engine, while on the contrary, the anchor and the subsequent take-up by a cued interactant could well be part of a student annotation project concerned with investigation of discourse patterns that cause unexpected disruptions to grammatical patterns.

Thanks to the active participation that the annotation of scenes entails, discourse analysis, which might otherwise be considered a rather dull activity, can be turned into a highly active and interactive process of discovery and reflection on descriptive models. Hopefully, the *House Corpus Project* will lead to corpus annotation projects suggested by students themselves. If so, we suspect they may well be directed towards a better understanding of the cultural models hidden in a TV series such as *House M.D.*, most obviously comparisons of expectations about medical services in different parts of the English-speaking world as reflected in answers to questions like *Did the patient lie?* and *Did another doctor screw up?* constantly foregrounded in the *House M.D.* series. Whatever happens in the future, there is considerable satisfaction in knowing that, so far, teacher and student responses to the project have been more than positive.

A final thought relates to the research efforts being made to overcome the risk of corpus studies having little bearing on classroom activities owing to a disproportionate focus on word counts and frequency-based statistics. Our title, *Ain't that sweet,* is a song-like slogan encouraging investments in multimedia corpora that serve the interests of scholars and students by stimulating engagement with the complexities of English discourse. Hopefully, this slogan will work in the same way for others as it has for us.

Acknowledgements: Special thanks to Francesca Bianchi for her assistance and advice with the *House Corpus Project*, in particular as regards engaging undergraduate students in pilot online and offline annotational activities.

Bionotes:
Davide Taibi is a researcher at the Institute for Educational Technology, Italian National Research Council and part-time lecturer at the University of Palermo. His research activities are mainly focused on: Mobile Learning, Semantic Web and Linked Data for education, Open Educational Resources and Learning Analytics.

Ivana Marenzi PhD is senior researcher at the L3S Research Center, Leibniz University of Hannover. Her main research area is Technology Enhanced Learning in support of collaborative and lifelong learning with a special focus on ties between technology and communication. She has published Multiliteracies and e-learning2.0, Peter Lang, Frankfurt, 2014.

Qazi Asim Ijaz Ahmad is a software engineer at the German National Library of Science and Technology (TIB), Hannover. His main work includes development of an open source research information software VIVO, text mining and citation and information extraction.

Authors' addresses: davide.taibi@itd.cnr.it; marenzi@l3s.de; asimijaz@live.com

References

Ackerley K. and Coccetta F. 2007, *Enriching language learning through a multimedia corpus*, in “ReCALL” 19 [3], pp. 351-370.

Balcikanli C. 2010, *Long live, YouTube: L2 stories about YouTube in language learning*, in Shafaei A., Nejati M. (eds.), *Proceedings of the 2009 International Online Language Conference (IOLC 2009)*”, Universal-Publishers, Boca Raton, Florida, pp. 91-96.

Baldry A.P. 2016, *Multisemiotic Transcriptions as Film Referencing Systems*, in Taylor C. (ed), *A Text of Many Colours – translating The West Wing*, Intralinea special issue. www.intralinea.org/specials/article/2195 (10.03.2018).

Berners-Lee T., Hendler J. and Lassila O. 2001, *The semantic web*, in “Scientific American” 284 [5], pp. 34-43.

Bianchi F. 2015, *Integrazione e Apprendimento: I prodotti cinetelevisivi come strumento didattico linguistico e culturale per il mediatore e il migrante*, in “Lingue e Linguaggi” 16, pp. 237-263.

Bianchi F. and Marenzi I. 2016, *Investigating student choices in performing higher-order comprehension tasks using TED talks in LearnWeb*, in “Lingue e Linguaggi” 19, pp. 23-40.

Butler C.S. 1997, *Repeated word combinations in spoken and written text: some implications for functional grammar*, in Butler C.S., Connolly J.H., Gatward R.A. and Vismans R.M. (eds.), *A Fund of Ideas: Recent Developments In Functional Grammar*, IFOTT, University of Amsterdam, Amsterdam, pp. 60-77.

Cheshire J. 1991, *Variation in the use of ain’t in an urban British English dialect*, in Trudgill P. and Chambers J. K. (eds.) *Dialects of English. Studies in Grammatical Variation*, Longman, London and New York, pp. 54-73.

Coccetta F. 2011, *Multimodal functional-notional*, in Frankenberg-Garcia A., Flowerdew L. and Aston G. (eds.), *New Trends in Corpora and Language Learning*, Continuum, London, pp. 121-138

Coccetta F. 2019, *Old Wine in new bottles. The case of the adjacency-pair framework revisited*, in “Lingue e Linguaggi” 29, pp. 407-424.

Crockford D. 2006, *RFC4627: The application/json media type for javascript object notation notation (json).* [Online]. http://tools.ietf.org/html/rfc4627 (22.10.2019).

Elk C.K. 2014, *Beyond mere listening comprehension: Using Ted Talks and metacognitive activities to encourage awareness of errors*, in “International Journal of Innovation in English Language Teaching and Research” 3 [2], pp. 215-246.

Ewerth R. and Freisleben B. 2004, *Video cut detection without thresholds*, in *Proceedings of 11th Workshop on Signals, Systems and Image Processing*, PTETiS, Poznan, Poland, pp. 227-230.

Flowerdew L. 2012, *Corpora and Language Education*, Palgrave Macmillan, Basingstoke.

Gormley C. and Tong Z. 2015, *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*, O'Reilly, Beijing.

Kohlschütter C., Fankhauser P. and Nejdl W. 2010, *Boilerplate detection using shallow text features*, in *Proceedings of the third ACM international conference on Web search and data mining (WSDM '10)*, ACM, New York, NY, USA, pp. 441-450.

Krashen S.D. 1982, *Principles and practice in second language acquisition*, Pergamon, Oxford.

Law L. 2015, *House M.D. Corpus Analysis: A Linguistic Intervention of Contemporary American English*, in Li, L., Mckeown, J. and Liu, L. (eds.), *Proceedings of AsiaLex 2015*

Hong Kong: Words, Dictionaries and Corpora: Innovations in reference science, The Hong Kong Polytechnic University, Hong Kong, pp. 230-249.

Lee D. and Swales J. 2006, *A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora*, in "English for specific purposes" 25 [1], pp. 56-75.

Lehmann J., Isele R., Jakob M., Jentzsch A., Kontokostas D., Mendes P.N. and Bizer C. 2015, *DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia*, in "Semantic Web" 6 [2], pp. 167-195.

Lesnov R.O. 2017, *Using videos in ESL listening achievement tests: Effects on difficulty*, in "Eurasian Journal of Applied Linguistics" 3 [1], pp. 67-91.

Loiacono A. and Tursi F., *this volume*.

McEnery T., Baker P. and Cheepen C. 2002, *Lexis, indirectness and politeness in operator calls*, in "Language and Computers" 36, pp. 53-70.

Mendes P.N., Jakob M., García-Silva A. and Bizer C. 2011, *DBpedia spotlight: shedding light on the web of documents*, in *Proceedings of the 7th international conference on semantic systems*. https://dl.acm.org/citation.cfm?id=2063519&dl=ACM&coll=DL (22.10.2019).

Park S.M. and Cha K. 2013, *Pre-service teachers' perspectives on a blended listening course using Ted Talks*, in "Multimedia-Assisted Language Learning" 16 [2], pp. 93-116.

Richards J.C. 2015, *The changing face of language learning: Learning beyond the classroom*, in "RELC Journal" 46 [1], pp. 5-22.

Salway A. 2007, *A corpus-based analysis of audio description*, in Orero, P. and Remael, A. (eds.) *Media for all: Subtitling for the deaf, audio description and sign language*, Rodopi, Amsterdam, pp. 151-174.

Sinclair J. 1991, *Corpus, Concordance, Collocation*, Oxford University Press, Oxford.

Takaesu A. 2017, *Ted Talks as an Extensive Listening Resource for EAP Students5*, in Kimura K. and Middlecamp J. (eds.), *Asian-Focused ELT Research and Practice: Voices from the Far Edge*, IDP Education Cambodia, Phnom Penh, pp.108-126.

Tottie G. and Hoffmann S. 2006, *Tag questions in British and American English*, in "Journal of English Linguistics" 34 [4], pp. 283-311.

Wagner E. 2010, *The effect of the use of video texts on ESL listening test-taker performance*, in "Language Testing*"* 27 [4], pp. 493-513.

Watkins J. and Wilkins M. 2011, *Using YouTube in the EFL classroom*, in "Language Education in Asia" 2 [1], pp. 113-119.

Website references

Further information on the tools mentioned in this article

Boilerpipe	https://code.google.com/archive/p/boilerpipe/ (12.11.2017).
DailyMotion	https://www.dailymotion.com (05.04.2018).
Dbpedia Spotlight	https://github.com/dbpedia-spotlight/dbpedia-spotlight (10.09.2016).
Dbpedia	http://wiki.dbpedia.org/ (10.09.2016).
Elasticsearch	https://www.elastic.co (10.09.2016).
Jsoup Java	https://jsoup.org/apidocs/ (10.09.2016).
Lucene API	https://lucene.apache.org/ (10.09.2016).