Corpus Pragmatics (2021) 5:7–36
https://doi.org/10.1007/s41701-019-00072-x

ORIGINAL PAPER

Possibilities and Drawbacks of Using an Online Application for Semi-automatic Corpus Analysis to Investigate Discourse Markers and Alternative Fluency Variables

Christoph Wolk[1] · **Sandra Götz**[2] · **Katja Jäschke**[3]

Received: 28 June 2019 / Accepted: 9 December 2019 / Published online: 23 January 2020

Abstract

To overcome planning phases in spontaneous speech production, learners and native speakers use strategies such as (un)filled pauses, smallwords or discourse markers. Small scale studies in this vein have demonstrated that learners differ from native speakers in that they underuse smallwords and discourse markers, and rely on other fluency-enhancing strategies instead. In the present paper, we present a corpus-based study, which investigates fluency-enhancing strategies in four components of the *Louvain International Database of Spoken English Interlanguage* (LINDSEI; Gilquin et al. 2010), covering four learner English varieties, namely Spanish, German, Bulgarian and Japanese. We investigate 216 different fluencemes (i.e. fluency-enhancing features; Götz in Fluency in native and nonnative English speech, John Benjamins, Amsterdam, 2013) in 200 transcribed interviews with advanced learners of English. An online coding application, which was specially designed and programmed for this project, enables us to cover such a large amount of data. We report on the design, functionality and (dis-)advantages of the online application, the multilevel-coding system we implemented, and the methodological challenges we face in detail. We will also present the findings of one first pilot study where we exhibit considerable variation between and within learners of particular native languages concerning fluenceme frequencies, while distributional patterns of fluencemes are rather similar across varieties.

Keywords Fluency · Discourse markers · Advanced spoken learner language · Online coding application

✉ Christoph Wolk
Christoph.B.Wolk@anglistik.uni-giessen.de

[1] University of Giessen, Room: B 403, Otto-Behaghel-Strasse 10 B, 35394 Giessen, Germany

[2] University of Giessen, Room: B 411, Otto-Behaghel-Strasse 10 B, 35394 Giessen, Germany

[3] University of Giessen, Room: B 408, Otto-Behaghel-Strasse 10 B, 35394 Giessen, Germany

Introduction

One strategy for learners to overcome planning phases in spontaneous speech production is the use of discourse markers (e.g. *you know*, *like*, *well*; e.g. Hasselgren 2002). Previous (learner corpus-based) research on English as a foreign language (EFL) has shown that learners heavily underuse such discourse markers and instead prefer using alternative strategies, such as filled or unfilled pauses (e.g. De Cock 2000; Müller 2005; Gilquin 2008; Götz 2013). Studies in this vein, however, have mainly focused on single learner varieties, whereas contrastive analyses on learners' use of discourse markers and comparable fluency-enhancing strategies from different L1 family backgrounds have only rarely been undertaken.

In the present paper, we present a research project aiming to close this gap by investigating discourse markers, filled pauses, and other fluency-enhancing devices in four components of the *Louvain International Database of Spoken English Interlanguage* (LINDSEI; Gilquin et al. 2010), containing interviews with upper intermediate to advanced learners of English from four different language backgrounds (viz. German, Japanese, Bulgarian and Spanish). We plan to investigate (1) whether fluency is established in different ways, (2) whether particular discourse markers appear in different utterance positions, and (3) whether the use of fluency-enhancing strategies can be predicted by extra-linguistic parameters, such as age or gender.

This undertaking, however, presents numerous methodological challenges: in order to attain such wide coverage of varieties and variables, existing corpora need to be leveraged, and a convenient and fail-safe method needs to be chosen or developed, because the large amounts of text involved require a whole team of coders. In this paper, we will thus particularly focus on the methods we use to ensure consistent, high quality annotations. We will mainly discuss the process of data coding, including the development of an online application that automatically extracts discourse markers and alternative fluency-enhancing strategies, or "fluencemes" (Götz 2013), from these learner corpora and displays them in their communicative context to facilitate the disambiguation of their use. We will also discuss the challenges of maintaining consistency and quality across the coding and annotation process. We will then present some first findings derived from a pilot study on discourse markers in the four learner varieties under scrutiny and conclude this paper with a summary and an outlook.

Discourse Markers and Fluency in English as Foreign Language

Discourse Markers as Fluencemes

Discourse markers were long associated negatively with a speaker's "unclear thinking, lack of confidence, [or even their] inadequate social skills" (Crystal 1988: 47), but in the past three decades their significance has been reviewed and research into discourse markers in both native and non-native speech has

flourished. Various studies have documented a variety of forms they can take and communicative functions they can fulfill. Erman (1987: 2) notices the huge terminological diversity that has been attributed to what we call discourse markers in the scope of this project (a few labels being pragmatic markers, verbal fillers, fumbles, softeners, pause-fillers, hesitation-markers, cajolers, pragmatic particles or turn holders). Despite the fact that discourse markers typically do not carry grammatical information, their use has still been found to be rule-governed (Erman 1987), so that using discourse markers has consequently been found to contribute to a speaker being perceived as "typically native-like", "natural-sounding" and "idiomatic" (De Cock 2000: 52). There is a considerable amount of studies that have investigated the phonological, syntactic, semantic, pragmatic and discourse-specific features of discourse markers in detail (e.g. Erman 1986; Schiffrin 1987; Lenk 1998; Aijmer 2002; Müller 2005; Crible 2018, to name but a few); sometimes they even fulfill various functions at the same time (e.g. using *well* at the beginning of an utterance to take a turn, while at the same time functioning as a planning device for the utterance to come).

The relationship of discourse markers and fluency has only recently entered into the spotlight. We take a similar approach and focus on the use of discourse markers only as strategies to overcome planning phases, resulting in a contribution to a speakers' productive fluency. However, we acknowledge that there is no obvious clear-cut functional distinction, so that some of the discourse markers we investigate will be polyfunctional and naturally also carry discourse-pragmatic functions in addition to their fluency-enhancing function (e.g. Hasselgren 2002; Götz 2013; Dumont 2018). In line with Shriberg (1994), who suggests an analogy between discourse markers and filled pauses or "fillers" (cf. also Swerts 1998; Pawley and Syder 2000; Götz 2013; Tottie 2011, 2015), we consider discourse markers to be only one of several equivalent strategies a speaker can select to overcome planning phases.[1] We consider all these strategies to be "fluencemes" in line with Götz (2013: 8–9). Fluencemes are thus abstract variables that contribute to a speaker's fluency, whatever their concrete realization may be. We would like to propose that a wide variety of fluencemes can be used interchangeably as devices to enhance productive fluency, because they can be used to overcome planning difficulties in all positions in the utterance. Alternative fluencemes are, among others, unfilled pauses (i.e. silences), filled pauses (e.g. *er*, *erm*, *eh*, *ehm*, to list but a few of their formal realizations) and vagueness markers (or "smallwords"; cf. Hasselgren 2002) that do not contribute to the content of an utterance (e.g. *sort of/sorta*, *kind of/kinda*, *quite*; cf. Hasselgren 2002).[2] Two further fluencemes we investigate with our application are repeats (e.g. *I I think so*) and incomplete utterances.[3] However, unlike previous research that

[1] As pointed out by one of the reviewers, we acknowledge different approaches towards discourse markers differentiating between discourse markers that increase and those that decrease fluency [for a discussion, see, for example, Crible (2018)].

[2] Please note that although we borrow the term "smallword" from Hasselgren (2002) we use it in a slightly different way by only including vagueness markers, while she also includes discourse markers, which we investigate as a separate category.

[3] The different fluencemes will be described in detail in Sect. 3 of this paper when we describe the data extraction and coding procedure.

investigates closed and often short lists of discourse markers either because of their frequency or their relevance (e.g. Müller 2005; Denke 2009; Götz 2013), in line with Crible's (2018) proposal, we shall first take a bottom-up approach to identify as many types of fluencemes as feasible that are used in a fluency-enhancing function, before we extract these in a top-down manner (see Sect. 3).

Fluencemes in EFL

In English as a native language (ENL), discourse markers occur with considerably high frequencies (cf. Biber et al. 1999). Very much in line with research on alternative planning strategies (such as filled or unfilled pauses), in native speech they mainly occur at the beginning of utterances (e.g. Biber et al. 1999: 1086) or at utterance boundaries (Erman 1986: 132), although they are generally independent of clause structure (Schiffrin 1987).

When speaking in a foreign language (EFL), the planning pressure of formulating an utterance is naturally higher than in ENL. To overcome these planning phases, discourse markers can serve as elegant fillers in comparison to alternative planning strategies while, at the same time, their use can increase the length of a speech run (and thus, a learner's productive fluency). Previous research on discourse markers in EFL of learners from different language backgrounds has documented a significant relative underuse regarding both their overall frequency, as well as regarding the use of different types of discourse markers by the same speakers (e.g. Erman 1987; De Cock 2000; Hasselgren 2002; Müller 2005; Gilquin 2008; Mukherjee 2009; Götz 2013; Dumont 2018). This heavy underuse might stem from the fact that an explicit teaching of discourse markers as a fluency-enhancing strategy has not been systematically integrated into EFL textbooks (e.g. Römer 2005) or classrooms. Studies investigating learners' use of discourse markers in comparison to alternative fluencemes have shown that learners exhibit a preference for alternative strategies over discourse markers and, compared to native speakers, use higher numbers of filled and unfilled pauses and repeats respectively (e.g. Gilquin 2008, Götz 2013; Dumont 2018). Additionally, fluency in general has been demonstrated to be correlated with learners' communicative behavior in their L1 (e.g. Peltonen 2018). As a result, learners sometimes even use discourse markers from their L1 instead of target-like ones. This has been shown, for instance, for German learners of English making use of the German particles *ach* or *ja*, or French learners of English using *hein* or *allez* when speaking English (e.g. Gilquin 2008; De Cock 2019).

Methodologically, in previous corpus-based research on fluency, discourse markers and alternative fluencemes have mainly been investigated in a 'top-down'-manner, i.e., a (typically small) set of discourse markers was identified and investigated in the learner corpus and manually disambiguated (e.g. Gilquin 2008; Hasselgren 2002; Müller 2005; Götz 2013; Dumont 2018; etc.), before being analyzed in isolation. More holistic approaches towards fluency in EFL that analyze a broader set of fluencemes (e.g. Götz 2013; Dumont 2018) are only able to focus on small datasets including only one learner variety. Our main aim is to present an online application that combines features of various data extraction tools so that many coders can work

on the data simultaneously, while at the same time including a more convenient and clean interface (compared to, e.g., classic concordancers), thus aiming at minimizing error-rates. In the present paper, we therefore focus particularly on the methodological aspects of this endeavor. In the remainder of this paper, we will describe the development of an application that will allow us to automatically extract fluencemes from the corpora and will show them in their communicative context, so that they can be disambiguated and labeled more conveniently. We will describe the app development, data extraction and coding procedure in detail in the following.

App Development, Data Extraction and Coding Procedure

Our analyses are based on four components of the *Louvain International Database of Spoken English Interlanguage* (LINDSEI; Gilquin et al. 2010), namely the German component (LINDSEI-GE), the Spanish component (LINDSEI-SP), the Japanese component (LINDSEI-JP) and the Bulgarian component (LINDSEI-BG). Each subcorpus of LINDSEI comprises 50 orthographically transcribed interviews that were held with English majors in their 3rd or 4th year of study at university. All interviews were conducted according to the same criteria, starting with a short monologic part (e.g. about a stay abroad or a film/play the learners liked), followed by a dialogic part about everyday topics and a retelling of a picture story, which makes LINDSEI an ideal resource for a contrastive (interlanguage) analyses (Granger 1996, 2015) of learners with different L1s. The components are of uneven lengths; counting only material uttered by the learners but including pauses, GE is the largest at about 100,000 words, JP the smallest at around 40,000, and both BG and SP are in the middle at 70,000 words each.

As fluency cannot be directly measured in this corpus, we extract and analyze several fluencemes uttered by the same speaker to quantify their (productive) fluency. On the technical side, we use a set of software tools to prepare and conduct the data extraction and annotation procedure. All of them were developed by the researchers using the statistical programming language R (R Core Team 2018) and consist of (1) several scripts to process the original data and identify fluencemes, (2) a web application that displays the fluencemes in their local context to facilitate efficient and accurate disambiguation, and (3) analysis scripts that operate on the coded data. We will discuss this web app in more detail below.

The semi-automated coding procedure of these fluencemes consists of several steps which will be outlined in the following. First, the potential fluenceme instances have to be identified. Our first set of tools prepares the corpora by parsing the corpus format and identifying the material where both speakers are overlapping, outputting the corpus to a suitable intermediate format. This is necessary to display the interviews in a more easily readable manner, and for automatically identifying some contextual factors, such as whether the individual token is at the start of a turn. The resulting files are then searched for instances of potential fluenceme uses, and the extracted list is uploaded to the web application, where the annotation team can interact with them. This corpus tool displays the fluencemes in the corpora and allows for accepting, rejecting, and commenting. The identified fluencemes are

manually disambiguated by a team of coders to determine whether e.g. the string *like* is a verb or a discourse marker. Finally, our analysis scripts operate on the final dataset and determine, for example, the total frequency of a particular fluenceme type in a file and how they are distributed throughout the interview. We will describe the individual steps in more detail below; the general procedure, however, is straightforward: identify search terms > extract candidates > two or more rounds of coding and correction > analysis.

Relevant Fluencemes

The first step of the coding procedure is to identify as many fluencemes as possible which appear in the corpora we are investigating. We constructed an initial list of fluencemes based on the literature (Götz 2013; Hasselgren 2002 among others), which we extended using generated word lists and through manual inspection of the corpus files. The list of fluencemes was continuously updated throughout the coding process. Our work on these corpora, as well as parallel research on other corpora, has led to the identification of 60 different fluencemes (counting different lexical realizations, but not variant spellings) that could plausibly appear in these LINDSEI components. These can be classified into several types and subtypes, i.e. discourse markers (e.g. *like*, *well*, *you know*, but also discourse markers from the speakers' L1s, such as *ach*, *ja*, *eto*, etc.), smallwords (e.g. *sort of*/*sorta, kind of*/*kinda*), filled pauses (e.g. *ehm*, *uh*, *uhm*), pauses and other features which are signs of disfluencies, such as repeats (e.g. *you know you know*) and incomplete words (e.g. *the fir=first of January*). An overview of the fluency types, some of the subtypes and examples can be found in Table 1, and an exhaustive list in the appendix (Tables 3 and 4). Some of the features, in particular both types of pauses, are explicitly annotated in the corpus. We rely on the accuracy of the transcription for these, although we did check for common realizations of filled pauses that were not explicitly annotated.

Repeats can be difficult to identify automatically, as in principle, any word, or sequence of words, can be repeated. Technological solutions to this are quite straightforward, as one can simply check whether two words are equal, but there is a further complication. In between the repetitions, or even within the repeated constituent, other material may appear that would not invalidate it as a repeat. The clearest example is an unfilled pause (.): if *the the* counts as a repeat, then surely *the . the* should as well, and so should *the . well the*. We therefore use a multi-stage process to select as many cases as possible for manual disambiguation. First, all other fluenceme types are marked in the corpus. Then, each word is compared to the following word, marking it as a repeat if the two words are equal. This step handles the typical case, and will also identify repeated fluencemes. Then the same process is repeated, but all other potential fluencemes are excluded first. This step handles instances where other fluencemes intervene within or between repeated material. We then iterate this process to consider not only individual words but also longer

Table 1 Types and subtypes of fluencemes (selected examples)

Type	Fluenceme	Example
Discourse markers (DM)	*I don't know*	"[…] and what I liked in Washington is . I saw I think it was (er) I don't know a street and there was a […]"; (LINDSEI-BG 004)
	like	"[…] but when they are dancing and drinking the ca= in like in the bar […]"; (LINDSEI-JP 003)
	okay	"[…] (mhm) (mhm) . okay . I'd like to go on talking about films but actually I <starts laughing> […]"; (LINDSEI-GE 031)
	well	"*[…]* (mm) .. well last year I spent my holidays there and we made a tour . through several towns[…]"; (LINDSEI-GE 044)
	you know	"[…] I I met lots of people and . you know . still I'm still in touch with some of them so […]"; (LINDSEI-SP 002)
	you see	"[…] it's not that it's just sad it's (erm) . you see I don't know . uneasiness . a feeling I don't know . quite . uncomfortable […]"; (LINDSEI-SP 005)
Smallwords (SW)	*kind of*	"[…] yes I think because (eh) yes it was kind of … (erm) … a thing to (eh) think about […]"; (LINDSEI-BG 046)
	stuff like that	"[…] I mean friends of mine had children who said oh no you can clean up because my (mm) pays you fifty pounds a week and stuff like that so […]"; (LINDSEI-GE 009)
	sort of	"*[…]* so well . he must be thinking of . she's sort of silly […]"; (LINDSEI-SP 006)
Filled Pauses (FP):	*eh*	"*[…]* (eh) yes (eh) . I mean they were .. (eh) … they were wearing short skirts very short skirts very open (eh) .. tee-shirts and blouses and everything […]"; (LINDSEI-BG 016)
Unfilled Pause	...	"*[…]* (eh) … (mm) I. I'm a student from Sho= Showa Women University […]"; (LINDSEI-JP 001)
Other	repeats	"*[…]*and I left . I I took the . the flight (em) . on Christmas Day[…]"; (LINDSEI-BG 001)
	incompletes	"[…]and this was also kind of factor that in= influenced my great impression […]"; (LINDSEI-BG 047)

A full list of ENL and EFL fluencemes, relevant for this paper, is provided in the appendix

sequences of up to 11 words. Incomplete words, however, are never considered to be a repeat, even if there is a partial match with the following word.

In the case of the EFL corpora, we additionally include fluencemes which originate in the speakers' native languages. Foreign (i.e., non-English) words are tagged (as *<foreign>)* in the LINDSEI corpora. Thus, we were able to identify potential foreign fluencemes by generating bigrams of the tag *<foreign>* and the adjacent words in *AntConc* (Anthony 2014). Whether the detected foreign words are fluencemes or not, is manually checked by researchers with sufficient language proficiency in the pertinent language or by native speaker informants.[4] Some of the foreign fluencemes we identified and examples of their use are listed in Table 2.

In LINDSEI Bulgaria, foreign fluencemes were rare. We found only two instances in all available files, namely the Bulgarian conjunctions *ami* and *mi mai*. The other corpora exhibited a higher frequency of foreign fluencemes, up to 84 in LINDSEI-JP. In German *also* and *ja* can be used as fluencemes; note that German *also* does not correspond to the meaning of the English word *also*, but rather to the meaning of *well*.

The native and foreign fluencemes discussed in this section form the basis for our analysis. They are uploaded to the corpus tool together with the corpora. The corpus tool allows a semi-automated coding process that will be described in the following.

Corpus Tool

The corpus tool is an online application that has been specially designed and programmed for this project and facilitates the coding procedure in several ways. In this discussion, we will focus on the use of this software as part of our coding process, and refrain from discussing the technical details. Like the rest of our software tools, it was developed in R, and leveraged the web application framework *shiny*.[5] Using a web app as the main coding interface has several advantages.[6] As it centralizes data storage, everything is immediately accessible to other team members, without any need for manual version control to keep everyone up to date. As the annotation team works with the application interface and not the data files directly, there is little risk of data incompatibilities resulting from different software, or coding mistakes such as spelling errors. Also, as we shall demonstrate now, it can facilitate the correctness of manual disambiguation. The app displays the interviews, which are linearly transcribed in the corpora, as dialogues and thus enables the coders to read them easily. This is a considerable advantage compared to using unformatted text files, which contain irrelevant annotation and display overlapping speech sequences in sequence, not in parallel, making the conversation hard to follow. At present, we only leverage

[4] We would like to thank Prof. Tania Kuteva and Birgyl Nier for providing helpful insights into the foreign words we found in LINDSEI-BG.

[5] More information about shiny is available at https://shiny.rstudio.com. The source code for the interface is available from the first author on request.

[6] Note, however, that existing software packages, such as Praat or Exmaralda, also cover substantial parts of the feature set of our tool, such as a clear display of transcriptions.

Table 2 Examples of foreign fluencemes in EFL

Corpus	Fluenceme	English equivalent	Example
LINDSEI-BG	*ami*	*but*	"[…] well (erm) . there was: . the father of . well one of the girls: didn't have ami a good family at all […]"; (LINDSEI-BG 011)
	mi mai	*as if*	"[…] (erm) . mi mai . yeah […]"; (LINDSEI-BG 037)
LINDSEI-GE	*also*	*well*	"[…] (erm) also after my A levels and: . I I really liked it . I think I I'd never go to: to the . to America […]"; (LINDSEI-GE 006)
	ja	*yes*	"[…] ja . (erm) well I've made a . a survey . last year with a (erm) […]"; (LINDSEI-GE 017)
LINDSEI-JP	*etto*	Japanese filled pause	"[…] etto (er) I (er) my school name is Showa Women's University and that is in Sangenjaya […]"; (LINDSEI-JP 015)
	nandakke	*What is it? How can I say?*	"*[…] in America .. I saw many … (mm) nandakke .. shingo mushi (lit. "not observing/not paying attention to the traffic lights") […]*"; (LINSEI-JP 045)
LINDSEI-SP	*pues*	*so, now, well*	"[…] (eh) it's (eh) . I don't know . pues well the the the killer (mm) he […]"; (LINDSEI-SP 011)
	bueno	*well*	"[…] it was like . bueno I mean I n= . I never had any experience with handicapped people […]"; (LINDSEI-SP 014)

the interactive web application for disambiguation and inspection; all analyses of the results are done in the standard way (i.e. using spreadsheets or specific analysis scripts).

The tool highlights the potential fluencemes that were pre-selected from the corpora. Thus, the coders do not have to search manually or interact with the corpora directly. This eliminates any possibility of oversights, which would certainly be a danger given the high number of potential fluencemes and text files.

The tool's interface also permits a convenient way to add comments to individual tokens if necessary. The corpus tool's interface is displayed in Fig. 1.

All uploaded corpora and files are easily accessible for the coder in the sidebar (see Fig. 2). Fluencemes can be coded in the order in which they appear in the corpus, or filtered by type (i.e. discourse markers, foreign fluencemes etc.) or lexical realization (i.e. *like*, *you know*, *etto* etc.). The control elements are dynamically updated, so that if the user selects, for example, the type DM (discourse marker), only the different discourse markers will be displayed in the list of realizations, and only these will be available for inspection and coding. While the tool will show a reasonable amount of local context by default, the coder can adjust the window size by adjusting the sliders at the bottom (here ranging from position 96 to 158).

The right-hand side of the corpus tool's interface (Fig. 3) primarily shows a segment of the current dialogue, highlighting potential fluencemes and underlining the current candidate. Simple buttons are used for accepting or rejecting the candidate as well as postponing the decision. The graph below the dialogue window displays the distribution of fluencemes as colored boxes.

Coding Process

Our method can only detect string identity, not determine whether a particular instance actually functions as a fluenceme (e.g. *I was like really surprised*) or not (e.g. as a verb in *I like cats*). The corpus tool's results therefore have to be manually disambiguated by a team of coders.[7] This may seem like a simple decision in most cases, but can often become quite challenging. The extreme variability of spoken language, especially learner language, complicates the development of a transparent and reproducible coding system. We thus implemented a multi-level coding procedure, which includes several rounds of correction to ensure a consistent high coding quality. Here, first, each corpus is disambiguated by one coder. Afterwards, the first coder's judgments are checked by a second coder. Cases in which the first and second coder's judgements deviate from one another, which are unclear or raise general questions are checked by a researcher. Items which remain unclear after these three coding steps are discussed by the whole team consisting of five to seven coders and researchers and are decided by majority decision if necessary. We keep a log of all the items on which the coders disagree, so we are able to reassess any problematic

[7] The coders are student assistants who major in English. They were given a coding manual and a baseline was established through shared coding of sample files and subsequent discussion. Regular meetings were held during the coding process to allow for discussion between coders and researchers.

Fig. 1 Interface of the corpus tool

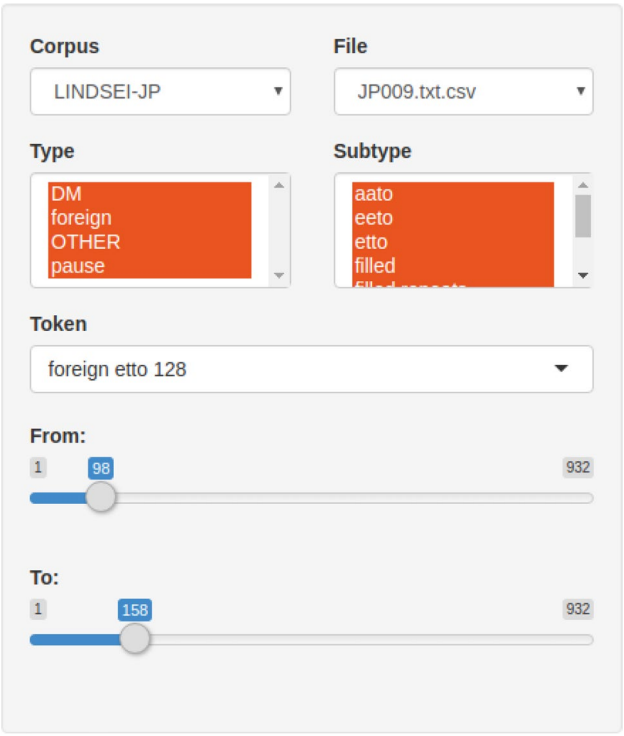

Fig. 2 Interface of the corpus tool: sidebar

case. Unfortunately, using the audio files to help with disambiguation is not feasible at the moment without severely reducing the number of files included in the analysis. Coders did, however, mark tokens as ambiguous, so that future projects can revisit these decisions.

Lastly, each corpus will go through a final correction round to make sure that all the decisions made throughout the coding process are consequently implemented. In the following, we will introduce the disambiguation process in some more detail and discuss some examples. Space prevents us from publishing our full coding manual, but it will of course be included with the future release of the full data set. The simplest category is unfilled pauses,[8] as they are automatically accepted as fluencemes.

Filled pauses (e.g. *uhm*, *eh, mhm*) cannot always be accepted as fluencemes (1a) as the same non-verbal sounds are also used for back channeling (1b), to answers questions, or as reactions to utterances (1c). In example (1c) it is unclear whether speaker A uses *mm* as a reaction to B's utterance to express agreement or comprehension, or whether it is indeed a strategy to enhance fluency. Unclear cases like this

[8] The corpora do not contain detailed information on pause lengths, only a rough classification. Research has shown that pause length can also have significant effects on learner fluency (e.g. Dumont 2018); unfortunately, taking this into account is not feasible in the scope of our project.

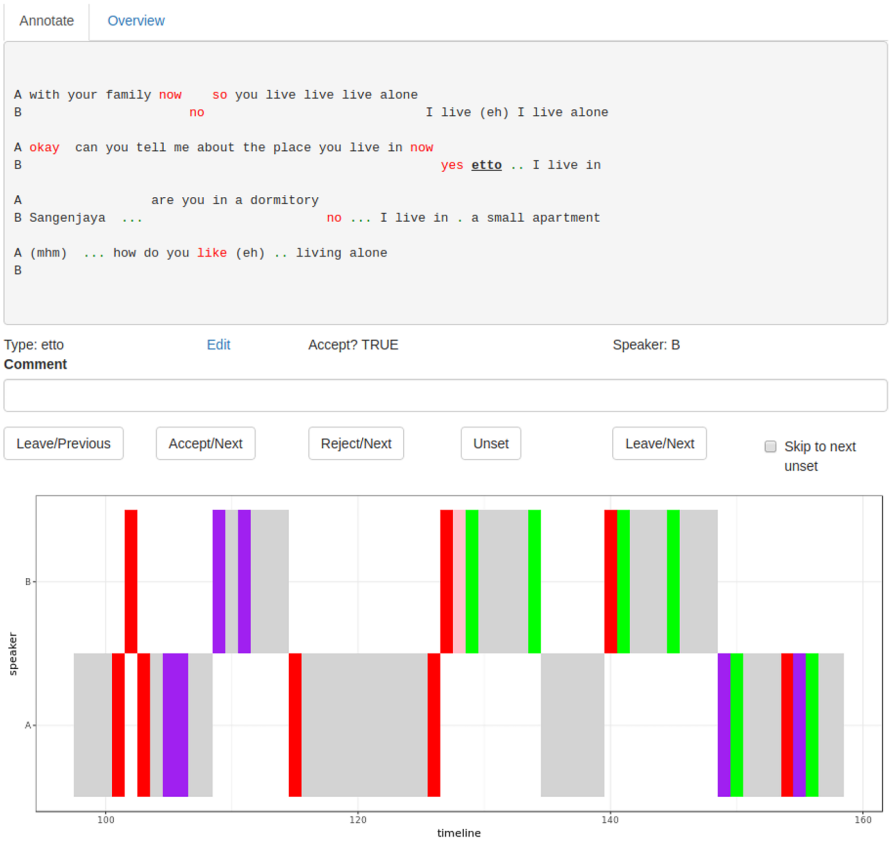

Fig. 3 Interface of the corpus tool: main coding interface

are categorically rejected. Cases in which the non-verbal sound is a clear reaction to the utterance are also rejected as they are not fluencemes, either.

(1) a. "[…] (mhm) Setagayais (mm) . almost is house <laughs> house building[…]" (LINDSEI-JP 011)

b. **A**: "(mm)"
B: "I am a Showa University student my major is English literature" "but my
A: "how did you get interested […]"
B: course is linguistics so I study a linguistics (mm)"
(LINDSEI-JP 011)

c. **B:** "[…] I feel it's kind of boring for me to live there so I prefer New York but
B: (erm) Los Angels"
A: "(mm) so would you like would you like to go back to
A: New York someday […]"
(LINDSEI-JP 016)

The general idea for disambiguation is quite straightforward: If a filled pause, discourse marker or smallword is uttered to enhance fluency, it should not carry (propositional or conceptual) meaning. Therefore, neither meaning nor grammatical

correctness of a given utterance should change if the fluenceme is omitted, as shown in example (2). In sentence (2a), *now* functions as a fluenceme and can be omitted without changing the meaning of the utterance. It could also be replaced by another fluenceme like *well* or a filled pause like *uhm.* This does not hold true in sentence (2b), in which *now* is part of the adverbial *right now* and could neither be omitted nor replaced by another fluenceme without changing the meaning of the sentence or making it ungrammatical. Thus, *now* in sentence (2a) is accepted as a fluenceme while *now* in sentence (2b) is rejected.

(2) a. "[…] that's great . . now . do you wanna go in the winter just to see what it's like […]"
(LINDSEI-BG 030)

b. "[…] only a little part of the movies that are going round right now are worth seeing you know […]"
(LINDSEI-BG 019)

While the examples in (2) are fairly straightforward, applying these principles can become quite complex. In the following, we will present selected examples to illustrate the challenges this task poses. In particular, we will focus on a case where our usual heuristics cannot be applied directly, namely potential fluencemes that occur in utterance final positions.

Whenever filled pauses, discourse markers and vagueness markers occur in utterance-final position, it is challenging to conclusively determine whether they can be considered part of a fluency-enhancing strategy and thus be classified as fluencemes (i.e. be accepted) or not. This problem arises due to several uncertainties.

Firstly, it is not always clear whether the speaker who used the utterance-final fluenceme ended the utterance deliberately or whether they were interrupted by another speaker. While a potential interruption does not necessarily pose a problem for classifying an item as a fluenceme (e.g. filled pauses are quite unproblematic), it can cause uncertainties when an utterance ends, for instance, in *so*. In the case of *so*, it is sometimes unclear whether the utterance was ended deliberately and *so* was used to link the utterance to an implicit continuation (in which case *so* would be classified as a fluenceme) or whether the speaker wanted to use *so* to express consequentiality and was interrupted (in which case *so* would not be considered a fluenceme). Interruptions are also relevant when repeats are under consideration. Whenever a repeated sequence is interrupted by another speaker, we do not classify it as a fluency-relevant repeat, as we assume that the speaker is repeating the word/sequence to hold their turn.

Whether a speaker was interrupted or not is especially difficult to determine, as the corpus transcriptions do not contain any information on intonation. Due to the large amount of data, we are not able to listen to the original recordings in every difficult case. To decide whether we are dealing with a fluenceme or not could, however, depend on knowing whether the speaker was interrupted, as outlined above.

Another example (3) is utterance-final *well,* which can be interpreted as sign of disfluency which speaker A uses to take over, but it is also plausible that *well* is used to link the utterance to content which is not verbalized, as has been argued for utterance-final *then* by Haselow (2011), who presents a detailed account of its pragmatic

functions. This case is unproblematic, even though it allows for different interpretations, since *well* would be classified as a fluenceme in both cases. Utterance-final *well* only poses categorization problems in rare cases, e.g., if the context suggests that the speaker could have intended to use *well* as an adverb, but ended the utterance (deliberately or not) and the syntactic dependencies remain unclear.

(3) **B**: "[…] they pretended to be and it was very very impressive"
A: "(mhm)"
B: "and well"
A: "what did you think of the food"
B: "<laughs> I I liked it a lot"
(LINDSEI-SP 037)

It is obvious that not all discourse markers which occur in utterance-final position can function as a link to a statement which is intentionally left unsaid, e.g. utterance-final *like*. These discourse markers therefore have to be treated differently in the disambiguation process.

We first attempted to classify utterance-final fluencemes into two categories, i.e. one category which contains fluencemes which are always accepted as fluencemes in utterance-final position (e.g. filled pauses) and a second category which contains items which are categorically rejected as fluencemes in utterance-final position. It turned out that such a categorization does not do justice to the complexity of the data, as many cases are too context-dependent to be treated in such strict categories We thus opted for a function-dependent categorization. While we consider discourse markers which function as a link to unuttered statements as fluencemes, we reject discourse markers or filled pauses which function as tag questions (e.g. *you know*, *right*, *eh*) as fluency-relevant items, because tag questions are used to elicit an answer or approval from an interlocutor, but not to bridge a fluency gap.

Potential fluencemes which are used in a quote (e.g. *He said erm …*, *He said well….*) are also somewhat problematic, as it is usually not clear whether the speaker is using the potential fluenceme to bridge a planning phase, or whether the original speaker used the fluenceme and the speaker quotes it. In the latter case, the items would not be of interest for our project. As we consider filled pauses as unlikely to be quoted, as they usually do not serve a pragmatic or content-bearing function in an utterance, we consider them fluencemes used by the speaker. Discourse markers could, however, have pragmatically shaped the original utterance in a way which makes them worth being quoted. We therefore do not consider them to be fluencemes.

Our general aim is to establish coding rules for as many recurrent problems as possible, as we have experienced that individual judgments made by different coders may vary greatly. We realize that these rules are debatable and do not suit each individual case. To ensure a high-quality, transparent and reproducible coding system across all coders at all times, we have decided to adopt a transparent and reproducible coding system rather than a system which is more flexible with respect to individual decisions but which may greatly depend on an individual coder's reading.

Feasibility Study

While the native-speaker corpora are still undergoing the coding process described in the previous section, the learner corpora are complete except for a small number of instances still awaiting ambiguity resolution. This allows us to present our first findings concerning variation between learners with different native languages here as a first taste of the kind of insights that such a large-scale corpus-linguistic investigation can provide. Let us begin by broadly summarizing the dataset. In total, our method extracted 104,878 potential fluencemes, of which 41,796 were unfilled pauses and were therefore automatically accepted. The remaining 64,082 tokens were manually disambiguated, and 53.7% were accepted in this process, yielding a total dataset of 76,229 fluenceme uses. As all the tokens were judged by at least two coders and required unanimity or a group decision, inter-rater agreement of the final data is virtually 100%. We can, however, estimate the degree of inter-rater agreement during the coding process by looking at the proportion of initial disagreements (excluding unfilled pauses, which are not manually coded). Depending on the learner group, these ranged from 1.6 (BG) to 5.7 (JP) percent. In the vast majority of cases, the initial disagreement was immediately resolved, as it resulted from a simple error by one of the coders. Only for a small fraction of tokens did both coders maintain different interpretations; for BG and JP this concerns about 0.3%, for GE 0.5%, and for SP approximately 0.8% of all tokens. This suggests that our coding process is very reliable. The numbers presented so far include the speech produced both by interviewees and by interviewers; we coded both to be able to test for mutual influence in the future. In the remainder of the paper, we will only consider the interviewees, who constitute 63,147 of the 76,229 tokens in our dataset.

Overall Distribution of Fluencemes Across Learner Corpora

Let us now consider the distribution of fluencemes in general. There is considerable variation between the different learner varieties, but as it turned out, the differences within learner varieties were even larger. This concerns primarily the Spanish learners of English, for whom material was collected at two universities,[9] Murcia and Madrid. These two groups exhibit very marked differences and will be separated in the following analysis, with the group from Murcia being labeled SP2.

Figure 4 displays the overall frequency distribution using boxplots, normalized to per 1000 words. Due to the vast differences in overall frequencies, the types were separated into three frequency bands, which vary in the scaling of the y-axis. Unsurprisingly, unfilled and filled pauses are by far the most frequent fluencemes. For unfilled pauses, we find a very high range for BG, less variation for JP but with a similar median, and much less variation with a higher baseline for GE. The two Spanish locations exhibit markedly different behavior from one another. While Madrid is relatively similar to GE, Murcia has exceptionally

[9] The data for Japanese learners of English was also collected at two universities, but one contributed only a small number of interviews, and cannot be reliably evaluated by itself.

low frequencies for this feature: the interquartile range does not overlap with any other learner variety, and the median is only a third of the closest variety. We hasten to add that unfilled pauses can be challenging to transcribe, and the observed differences may result from different practices with regard to how they are handled during the corpus compilation process. Filled pauses show the opposite pattern: BG, GE, JP, and Murcia are relatively similar, with elevated frequencies for JP, but Madrid has a markedly lower rate of usage. The interquartile range shows only a minor overlap, and the median is less than half of the closest value in other varieties.

The mid-frequency band consists of discourse markers and repeats. Discourse markers follow a pattern similar to unfilled pauses. Again, there is a very marked difference between Madrid, where usage rates are quite high, and Murcia, where they are very low. While BG and GE rarely use repeats, we find a slightly more frequent use in the other learner varieties.

In the low-frequency band, we have incomplete words, smallwords, and foreign fluencemes. Incomplete words show minor variation, with frequencies in GE being somewhat higher; again, Murcia is a complete outlier, where this feature is particularly frequent. Smallwords are only commonly used in the BG and GE corpus, with at least 50% of texts in the other corpora containing none at all. Foreign fluencemes are even more extreme; here, at least half of all texts in all corpora contain no instances. There are, however, individual speakers that do make use of them in all varieties, particularly in JP (23 speakers) and GE (13 speakers).

Discourse Markers Across Learner Corpora

Let us now focus more specifically on discourse markers. The overall frequency exhibited a relatively consistent pattern, except for Murcia. Does this also hold for individual discourse markers? To investigate this, we extract and display the three most frequent variants in each variety; because of overlaps between these lists, this results in six discourse markers. The results are presented in Fig. 5. *Well* (4a) is the most common discourse marker both overall and in BG and Madrid. It is also quite frequent in GE, but rare in Murcia and almost absent in JP. Japanese learners exhibit a clear preference for *so* (4b) instead, their most frequent discourse marker. It is, however, relatively rare in the other varieties, except for Madrid. In GE, *yeah* (4c) is the most frequent one, although the difference from *well* is minor. While it is the third entry in the overall frequency list, it is relatively rare elsewhere, again with the exception of Madrid. Our fourth entry is *I don't know* (4d), which is used most often in Madrid, and is almost absent in JP. *Like* (4e) is fifth on the list and already quite rare, but used most in Madrid. Finally, we have *I mean* (4f), which exhibits similar rates in BG, GE, and Madrid, but is quite rare in JP and Murcia.

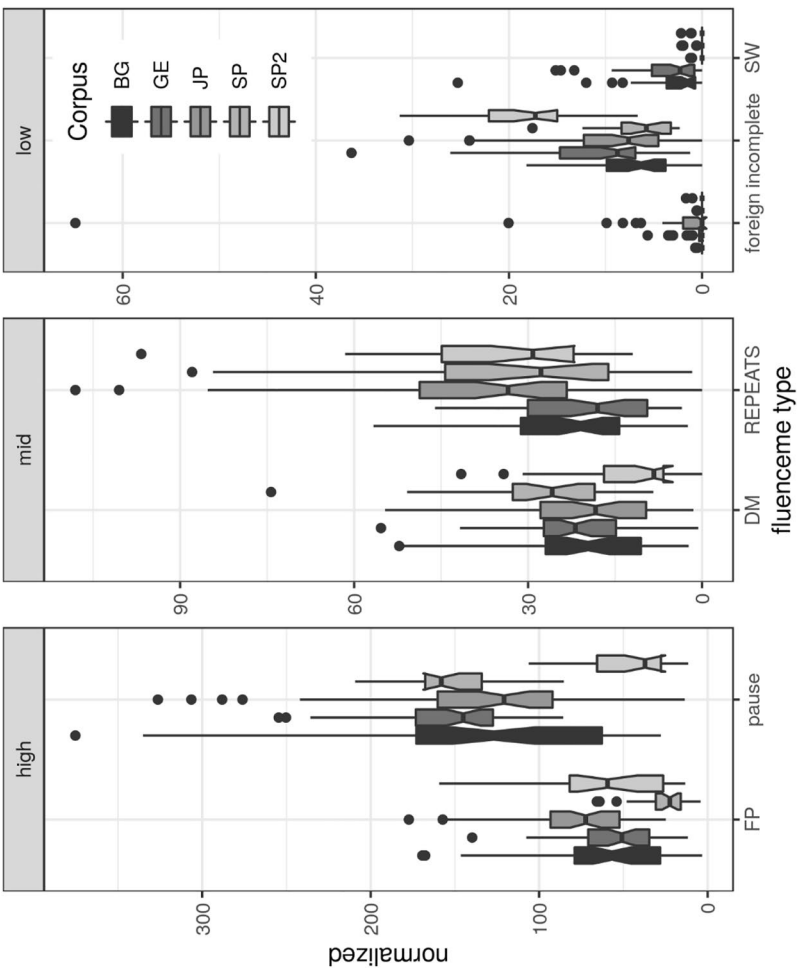

Fig. 4 Distribution of fluenceme types across learner groups, normalized to per 1000 words

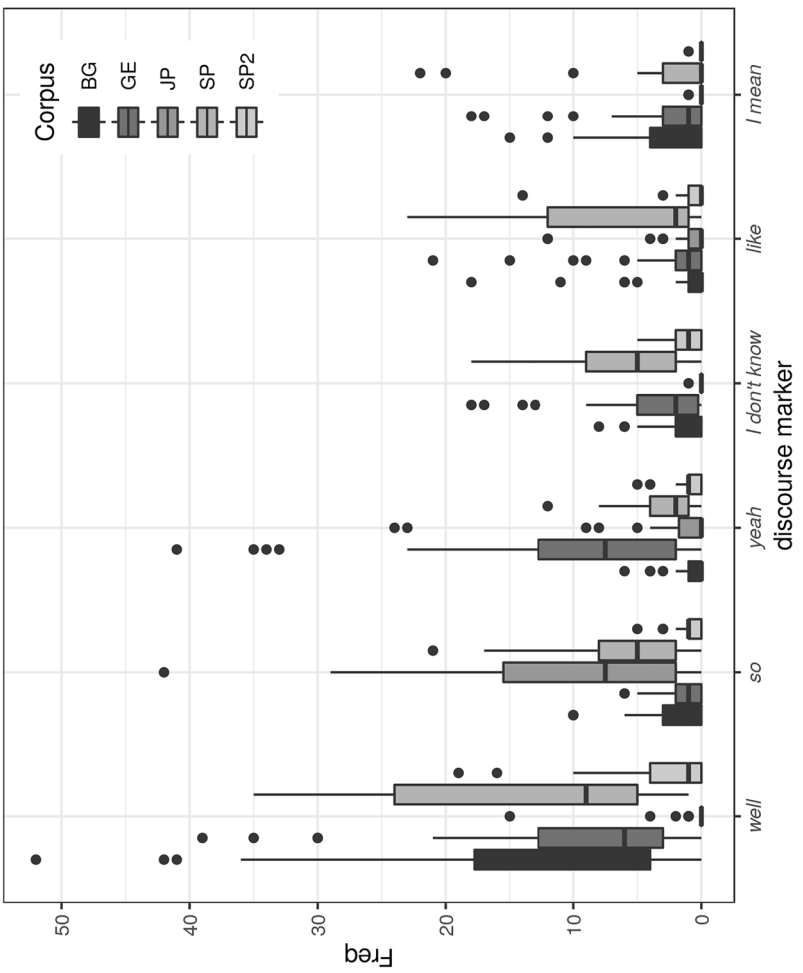

Fig. 5 Distribution of the most frequent discourse markers across learner groups, normalized to per 1000 words

(4) a. "Indians yeah and and and today's Mexicans they really don't mind these death images like sculls and .well those scary skin figures Christ figures everywhere" (LINDSEI-BG 001)

b. "[…] living myself is (eh) enjoyable (eh) but (eh) sometimes I feel lonely and I I remember my family *so* (mm) . I usu= I usually call call my family and usually I spea= (eh) I talk to my mother *so*. (mm) .. so . after I began to after I began to live by myself in Tokyo" (LINDSEI-JP 005)

c. "[…] last time I really . I really fussed about that topic and (erm) . well yeah . but on on Tuesday or whenever that was . (erm) . we started talking in a very .. well it was it was quite simple yeah and then […]" (LINDSEI-GE 010)

d. "it's wonderful because (eh) . it's always snowing . and . it's . I like it . and . I don't know." (LINDSEI-SP 023)

e. "yeah it's horrible because . like . I watch Friends a lot okay. so by the end I'm watching Friends I'm like . talking . kind of an American way" (LINDSEI-SP 007)

f. "and . (eh) yeah I found (mm) in Manchester very nice people and . so friendly I mean . you know . like . they just talk to you even if they not met you before and . oh it was great I mean" (LINDSEI-SP 014)

In short, there is considerable variation concerning the discourse markers used beyond the variation in overall frequency described in the previous section. BG, GE, and JP may contain discourse markers at overall quite similar rates, but they differ sharply in their preferences. BG, Murcia and JP have a strong preference for one discourse marker, namely *well* or *so*. GE has a preference for two discourse markers, *well* and *yeah*, and Madrid finally has a preference for *well*, but exhibits greater use of the other discourse markers as well. It should be kept in mind that there is also considerable variation within groups—taking BG as an example, we find that four speakers use *well* more than 20 times *ptw*, but four also do not use it at all.

Fluenceme Positioning Across Learner Corpora

Can the position of the fluencemes in the discourse help us understand these patterns? Our extraction process makes it possible to determine the distance to the last speaker change automatically, and therefore the beginning of the speaker's turn.[10] For this analysis, we count not only the first token uttered in a turn but, if that token is a fluenceme, also all other fluencemes that immediately follow it, so that a turn beginning with *uhm well* would count both as a filled pause and as a discourse marker. Bulgarian learners begin over 10% of their turns with a discourse marker; for the other learner groups, this value lies between around 4 and 7%. *Well* alone accounts for most of this high value for Bulgarian learners at almost 8% of turns, and this also accounts for the majority of total *well* uses there: only 45% of all instances of *well* are not turn-initial; (5a) contains an example showing both types. It is also the most common turn-initial discourse marker in the other learner varieties, except for JP, but they employ it much more rarely in this position than BG, namely

[10] For this analysis, we will consider backchanneling to be the start of a new turn, unless it is explicitly marked as an overlapping sequence in the corpus. We will revisit this and separate out the different contexts in future research.

in less than 1% of turns in JP and between 1 and 2% in the others. The association of the marker and the beginning of turns is also much weaker, with between 65 and 76% of uses happening in other positions. The high use of discourse markers, and particularly *well*, by Bulgarian learners therefore results from a strong preference for beginning turns this way. Japanese learners instead use their overall preferred discourse marker *so* to begin about 4% of turns, as in (5b), but there is no particular association to this position. The share of turn-initial uses is comparable to other learner varieties, with between 24 and 33% of all uses of *so* being turn-initial across groups. Spanish learners from Madrid exhibit the highest rate of use of turn-initial discourse markers except for BG, and those from Murcia the lowest; this is consistent with the overall frequency of discourse markers in those varieties. Regarding non-discourse marker fluencemes, learners from Germany are particularly likely to begin their turn with an unfilled pause[11] (see (5c)), namely in about 20% of turns, compared to 2–10% for the other varieties), and Japanese learners use filled pauses particularly often (as in (5d)), 17% for JP vs. 6–11% for the other groups). Both groups of Spanish learners are particularly likely to not begin their turn with a fluenceme at all, as in (5e), at almost 80% compared to around 65% for the other varieties.

(5) a. **A**: "[…] what are the memories of the past or what are the problems the family problems you mentioned"
B: "well (erm) . there was: . the father of . *well* one of the girls: didn't have ami a good family at all she had I mean her mother had problems with her father or: . something of the kind […]" (LINDSEI-BG 011)

b. **A**: "(uhu)"
B: "so . but they are very kindful kind very kind . and . (er) .. very friendly . and they . tr= tried to talk with talk with me"
(LINDSEI-JP 022)

c. **A**: "down near the Quay down the Circular Quay somewhere is it".
B: ". no it's not in the city at all it's: further outside it's Macquarie"
A: " Mac = Macquarie"
(LINDSEI-GE 015)

d. **A**: "what course are you in"
B: "*(eh)* I major in (eh) .. lin= linguistics"
(LINDSEI-JP 022)

e. **A**: "how different"
B: "because I think they . they don't cook . like here .. it's . only salads or .. or (mm) sandwiches . lots of sandwiches . sandwiches everyday <laughs>"
(LINDSEI-SP 012)

[11] In principle, unfilled pauses at turn boundaries present a challenge, as pauses between turns cannot necessarily be clearly assigned to either speaker. As the corpus consists of interviews with one speaker mostly asking questions, however, the choice tends to be clear. After a review of a sample of tokens, we have decided to follow the transcriptions and include these pauses in the analysis, but will consider a manual check of all relevant tokens in the future. We are grateful to an anonymous reviewer for pointing out this issue.

Fluencemes and Learning Context Variables

Finally, the large amount of data that can be collected using corpus-based methods allows us to relate the use of particular fluencemes to extralinguistic factors. We use linear regression on the normalized frequencies of individual learners for particular fluenceme categories, and stepwise model fitting to select the final model. The following variables were considered: the age and gender of the learner, the number of years of both school and university education in English, an indicator whether the student has had a stay abroad in an English-speaking country and the duration of that stay in months, and the corpus. Unfortunately, we cannot use these factors to illuminate the differences between the two groups of Spanish learners, as most variables are missing for the group from Murcia, who therefore had to be removed from this analysis. All significant effects are illustrated in Fig. 6.

For discourse marker frequency, only one variable survived the model-fitting process, namely the duration of university education. This factor is statistically significant ($p < .05$), and has the expected direction, with each year of university instruction increasing discourse marker frequency by two instances *ptw*. The model itself is significant ($p < .05$) but has a low explanatory power (adjusted $R^2 = 0.03$). For unfilled pauses, we find that both learner group and the duration of the stay abroad, if any, matter: all varieties use more unfilled pauses than BG, GE ($p < .01$) and SP ($p < .05$) significantly so. Furthermore, each month spent in an English-speaking country reduces the number of unfilled pauses by 1.3 *ptw* ($< .05$). The model is again significant, but the predictive power remains low ($p < .05$, $R^2 = 0.05$). Filled pauses have a similar pattern, except that the duration of a stay abroad does not matter; any stay reduces the average frequency by about 17 instances *ptw* ($p < .05$). Japanese learners use significantly more filled pauses ($p < .001$) than Bulgarian learners, and Spanish learners significantly fewer ($p < .05$). The model is again significant and has a much better fit than the previous models ($p < .001$, $R^2 = 0.21$). Finally, we have repeats, which Japanese and Spanish learners use more often than Bulgarian learners do. The number of years of school instruction in English and the speaker gender are also selected by the model-fitting process but are not individually significant. The model itself is, but has again relatively low model fit ($p < .001$, $R^2 = 0.13$).

Summary and Discussion

To summarize our findings: like in previous studies, our analysis has also shown that there is considerable variation both between and within groups of learners of particular native languages concerning fluenceme frequencies. Looking at comparable fluencemes, however, revealed that the distributional patterns of fluencemes are rather similar across varieties. All the learner groups show a preference for using filled and unfilled pauses followed by repeats, before we find them using discourse markers (target-like as well as foreign ones), smallwords and incomplete utterances. While this 'ranking', as it were, is similar to the fluenceme ranking in ENL (cf. Biber et al. 1999), the overall frequencies still seem to be much higher for filled and unfilled pauses and much lower for alternative strategies. Confirmation of this,

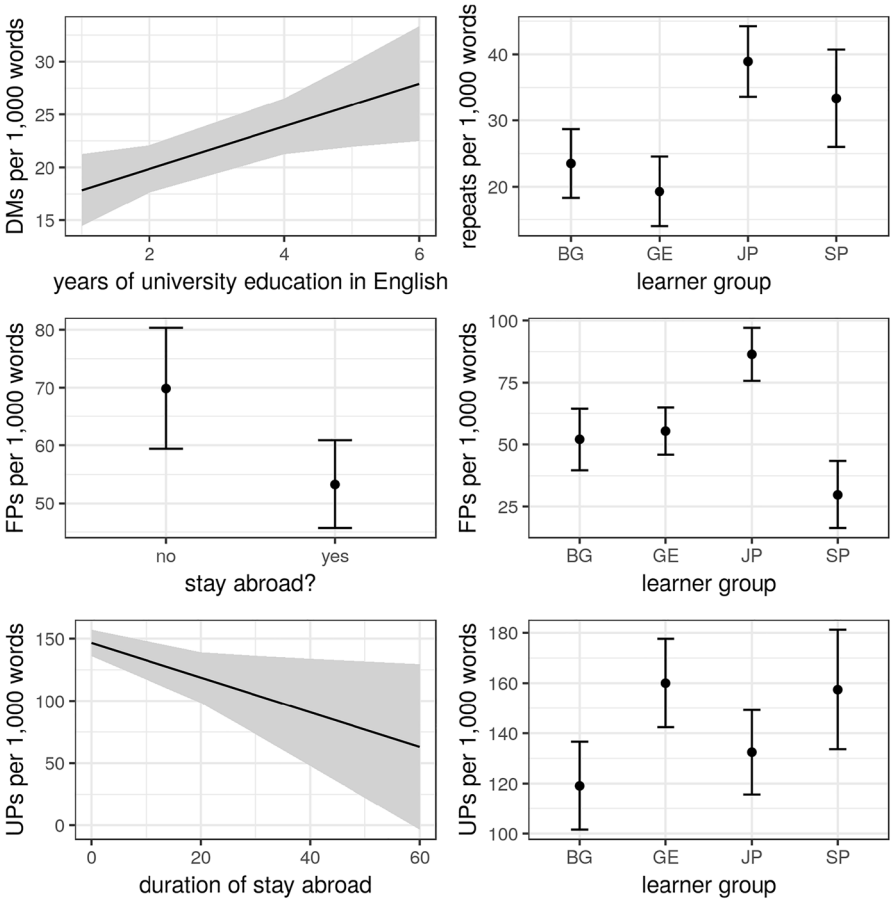

Fig. 6 Effect plots for the significant predictors of fluenceme frequency. Top row: discourse markers and repeats. Middle row: filled pauses. Bottom row: unfilled pauses

however, requires an ENL control corpus, which is still being annotated at the time of writing. Nevertheless, as far as overall frequencies are concerned, our pilot study seems to echo previous research, i.e. the learners use fluency-enhancing discourse markers and smallwords less frequently than alternative strategies that render the speech less fluent (i.e. filled and unfilled pauses) (cf. Sect. 2).

Zooming in on discourse marker use in particular, our pilot study indicates that a similar behavior across learner groups on the level of categories can hide clear distinctions on a lower level. In particular, there were considerable differences with regard to the use of discourse markers—*well* is generally the most common, but Bulgarian learners rely on it particularly heavily to start their turn, while Japanese learners avoid it and prefer the use of *so*. Other varieties use discourse markers more evenly, such as German learners, where *well* and *yeah* exhibit similar frequencies, and even more so Spanish learners from Madrid, who make relatively balanced use of a variety of discourse markers. The differences in fluenceme positioning across

learner varieties are quite noticeable and have only become possible to document and analyze thanks to our heavily computer-aided analysis. While our pilot study was only able to document these differences on a quantitative/distributional level, our follow-up research still needs to explain reasons for—and possible functional differences of—these different preferences across learner varieties.

We were also able to show that some language-external factors have an impact upon fluenceme frequency, including the duration of English instruction leading to an increased use of discourse markers and a stay abroad leading to a reduced use of filled and unfilled pauses (cf. similar findings in Gilquin 2016; Götz and Mukherjee 2018; Götz 2019). Since we can see significant positive correlations between both university instruction and a stay abroad and learner fluency, we would like to emphasize the language-pedagogical usefulness of both focused language practice courses at universities for English majors as well as promoting a stay abroad more rigorously, if a learner's goal is to improve their spoken fluency. Although the effect sizes of our models were fairly high, their explanatory power unfortunately remained rather low. This means that the learning context variables we investigated turned out to have a relevant effect but much of the variance in the data remains somewhat unexplained by these variables. We are looking forward to including further variables and conducting more fine-grained qualitative analyses to supplement the patterns that emerged through the quantitative analysis.

Conclusion and Outlook

In the present paper, we hope to have been able to highlight the benefits of using a corpus-based, semi-automatic quantitative approach to analyze discourse markers and further fluencemes in learner corpus data. This automatic approach enables us not only to analyze fluency more conveniently by (1) being able to analyze the distributional patterns of fluencemes in comparison to each other or select individual fluencemes, depending on the research question at hand, (2) being able to extract (a) the position of each fluenceme in the utterance and (b) the learning context variables for each speaker who uttered a particular fluenceme, while (3) a clearly-arranged interface shows fluencemes in their communicative context in the dialogue and thereby makes the extraction and disambiguation of a large amount of data quicker, more convenient and more accurate. However, we need to point out that this semi-automatic approach does not come without drawbacks: While the app enables the researcher to analyze their data more conveniently, the decisions and disambiguations still need to be made by a number of coders, including the selection of certain fluencemes over others. Therefore, when multi-faceted phenomena such as fluency are analyzed, the list of possible fluencemes that are included in the analysis needs to be permanently negotiated, updated and uploaded to the tool. The obvious drawback to this is the tremendous time and effort the coders need in preparing the lists of fluencemes and pre-investigating the data before the actual disambiguation procedure can begin. The list of possible fluencemes is also restricted to the ones that

can be found automatically, as other features that are relevant for a speaker's fluency (e.g. self-corrections, formulaic sequences) cannot be found with our approach. However, once the fluencemes are included in the tool, its flexibility is extremely powerful, as fluencemes, as well as interview contexts (monologue, dialogue, picture story retelling), can be included in or excluded from each data analysis in line with the research question at hand (e.g. one can only extract the category of unfilled pauses). Another drawback of using a corpus-based app is our reliance on the corpus transcripts alone. Analyzing the sound files as such, which would enable us to include intonation, prosody or pause lengths as additional fluency-relevant features, is not possible at the moment. One advantage of the approach that we have taken, however, is that our dataset can be easily and automatically linked back to the original corpus files, and our work can therefore be reused by any future projects that leverage the audio recordings of LINDSEI. We intend to release our data, once complete, for reuse by other researchers and projects, so that future work can build on this material and expand it to address such issues.

Finally, it goes without saying that developing, programming, maintaining and updating such a tool requires advanced programming skills, which are not necessarily available to everybody. If projects like these yield meaningful results, it might be useful to include the teaching of such skills in advanced corpus-linguistic modules at universities.

In the present paper, we have only been able to focus on the development of our methodology, its implementation and the findings of a first feasibility study we conducted on the basis of four learner corpora. In the context of this study, we followed a strictly quantitative approach to fluencemes regarding their frequency, position and distribution. Further research in the context of our project will need to include functional and qualitative analyses of these investigated fluencemes in order to be able to better explain these differences. Here, especially the use of discourse markers has been limited to their fluency-enhancing function so far; further research will also need to investigate learners' use of discourse markers in their respective pragmatic functions to uncover their polyfunctionality more thoroughly.

There are many further avenues we would like to explore within this research project. First, we would like to compare the learner data with native speaker data in order to reveal those areas where advanced learners still deviate significantly from the native target norm and those where they have already approximated to it when it comes to fluency. This approach also allows us to investigate whether fluency is prone to interference from the learners' L1s or if the mechanisms underlying learner fluency are universal in nature. Future analyses have the potential of giving us a better starting point to make some useful language-pedagogical suggestions on how to improve learner fluency, possibly also on the basis of using natural native-speaker data using our app. In a next step, we will take an even broader perspective and compare fluency within different speech communities. In doing so, we are planning to compare the learners' fluency to fluency in speech communities where English is spoken as a second language (ESL), namely Sri Lanka, India and the Philippines, as well as several speech communities in which English is spoken as a native language

(ENL; i.e. Great Britain, Canada, Australia and New Zealand). This will help us to systematically assess if speakers of these three different types of English establish fluency in generally different ways (e.g. by using different fluencemes) and if necessary planning phases become fewer the more institutionalized the English language becomes within the speech community (thus, suggesting a decrease in planning phases both in frequency and density from EFL to ESL to ENL). Finally, we plan to take the speaker type out of the equation and will solely investigate if the use of fluency-enhancing strategies can be predicted across speaker types by extra-linguistic parameters alone, such as age or gender. On this exciting journey, we have only been able to present the first—and maybe most important—step, namely the development of our taxonomy, the programming of the app and the process of developing the coding guidelines that will allow us to handle these large amounts of data. Although we are at the very beginning of this research, we consider the findings from our first small-scale feasibility study very promising and are convinced that special-purpose web applications will allow research on large-scale datasets in a shorter time and will thus offer multiple new options in linguistic corpus research.

Acknowledgements Open Access funding provided by Projekt DEAL. We would like to gratefully acknowledge that this research project has been generously funded by the *German Research Foundation* (DFG, Reference Numbers GO 1760/4-1 and WO 2224/1-1) as part of a larger project on "Fluency in ENL, ESL and EFL: A contrastive corpus-based study of English as a first, second, and foreign language". We would also like to thank our student assistants Lara Möller, Karola Schmidt, Hannah Vehrs and Daniel Walker for their invaluable help with the data coding and their patience with long discussions over fluencemes. We are grateful to two anonymous reviewers for their thorough and helpful comments. All remaining errors and infelicities are, however, our responsibility alone.

Compliance with Ethical Standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix

See Tables 3 and 4.

Table 3 Exhaustive list of all non-foreign fluencemes

Type	Fluenceme	Orthographic variants (if applicable)
DM (discourse marker)	*I don't know*	
DM	*like*	
DM	*okay*	*okay, ok*
DM	*right*	
DM	*alright*	*alright, allright*
DM	*well*	
DM	*you know*	
DM	*you see*	
DM	*I mean*	
DM	*know what I mean*	
DM	*you know what I mean*	
DM	*do you know what I mean*	
DM	*anyway*	
DM	*yeah*	
DM	*oh*	
DM	*or something*	
DM	*actually*	
DM	*anyhow*	
DM	*basically*	
DM	*now*	
DM	*let's see*	
DM	*so*	
DM	*no*	
DM	*just*	
DM	*nah*	
DM	*wayne*	
DM	*aye*	
SW (smallword)	*kind of*	*kind of, kinda*
SW	*quite*	
SW	*stuff like that*	
SW	*thing like that*	*thing like that, things like that*
SW	*in a way*	
SW	*sort of*	*sort of, sorta*
FP (filled pause)		*(many variants, explicitly annotated in LINDSEI)*
Unfilled pause		*(many variants, explicitly annotated in LINDSEI)*
Incomplete		*(many variants, explicitly annotated in LINDSEI)*
Repeat		*(many variants, extracted automatically)*

Table 4 Exhaustive list of all fluencemes classified as foreign

LINDSEI Subcorpus	Fluenceme	Orthographic variants (if applicable)	English equivalent
BUL	*ami*		*but*
BUL	*mi mai*		*as if*
GER	*ja*	*ja, ja:*	*yes*
GER	*ach*		*oh?*
GER	*uff*		*puh?*
GER	*also*		*well*
GER	*ah ja*		*oh yeah*
GER	*boah*		*wow?*
GER	*ach so*		*oh right*
GER	*naja*	*naja, na ja*	*well*
GER	*genau*		*right, exactly*
GER	*joa*		*yeah*
GER	*und*		*and*
JAP	*etto*	*untto, eto, eeto, mtto, ermtto, uuntoo, mmto, unto, unttoo, nto:, aato*	*Japanese filled pause*
JAP	*nandakkena*	*nandakkena, nandakke*	*What is is? How can I say?*
JAP	*wakannai*		*I don`t know*
JAP	*hai*		*yes*
JAP	*ja nakutte*		*it`s not like that*
JAP	*nanndaro*		*What is it? What could it be?*
JAP	*nannteiuu*		*What can I say? What is it called?*
JAP	*atto chigau*		*(Oh) That`s different.*
SPA	*pues*		*so, now, well*
SPA	*bueno*		*well*

References

Aijmer, K. (2002). *English discourse particles: Evidence from a corpus*. Amsterdam: John Benjamins.

Anthony, L. (2014). AntConc (Version 3.4.4w, Windows). Tokyo: Waseda University. Retrieved October 23, 2017 from http://www.laurenceanthony.net/.

Biber, D., Johannson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson Education.

Crible, A. (2018). *Discourse markers and (dis)fluency. Forms and functions across languages and registers*. Amsterdam: John Benjamins.

Crystal, D. (1988). Another look at, *well*, *you know*…. *English Today, 4*(1), 47–49.

De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair, & M. Hundt (Eds.), *Corpus linguistics and linguistic theory*. Papers from the twentieth international conference on English language research on computerized corpora (ICAME 20) (pp. 51–68). Amsterdam: Rodopi.

De Cock, S. (2019). Foreign words in EFL learner interviewee speech: Lending learners a productive fluency helping hand? In L. Degand, G. Gilquin, L. Meurant, & A. C. Simon (Eds.), *Fluency and disfluency across languages and language varieties*. Louvain-la-neuve: Presses Universitaires de Louvain.

Denke, A. (2009). *Nativelike performance. Pragmatic markers, repair and repetition in native and non-native English speech*. Saarbrücken: Verlag Dr. Müller.

Dumont, A. (2018). *Fluency and disfluency: A corpus study of non-native and native speaker (dis)fluency profiles*. PhD dissertation, Université catholique de Louvain.

Erman, B. (1986). Some pragmatic expressions in English conversation. In G. Tottie & I. Bäcklund (Eds.), *English speech and writing: A symposium* (pp. 131–147). Stockholm: Almqvist & Wiksell.

Erman, B. (1987). *Pragmatic expressions in English*. Stockholm: Almqvist & Wiksell.

Gilquin, G. (2008). Hesitation markers among EFL learners: Pragmatic deficiency or difference? In J. Romero-Trillo (Ed.), *Pragmatics and corpus linguistics: A mutualistic entente* (pp. 119–149). Berlin: Mouton de Gruyter.

Gilquin, G. (2016). Discourse markers in L2 English: From classroom to naturalistic input. In O. Timofeeva, A Ch. Gardner, A. Honkapohja, & S. Chevalier (Eds.), *New approaches to English linguistics: Building bridges* (pp. 213–249). Amsterdam: John Benjamins.

Gilquin, G., De Cock, S., & Granger, S. (2010). *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Götz, S. (2013). *Fluency in native and nonnative English speech*. Amsterdam: John Benjamins.

Götz, S. (2019). Do learning context variables have an effect on learners' (dis)fluency? Language-specific vs. universal patterns in advanced learners' use of filled pauses. In L. Degand, G. Gilquin, L. Meurant, & A. C. Simon (Eds.), *Fluency and disfluency across languages and language varieties* (pp. 177–196). Louvain-la-neuve: Presses Universitaires de Louvain.

Götz, S., & Mukherjee, J. (2018). The effect of the study abroad variable in spoken learner language: A pseudo-longitudinal study on spoken German learner English. In V. Brezina & L. Flowerdew (Eds.), *Learner corpus research: New perspectives and applications* (pp. 47–65). London: Bloomsbury.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Text-based cross-linguistic studies* (pp. 37–51). Lund: Lund University Press.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research, 1*(1), 7–24.

Haselow, A. (2011). Discourse marker and modal particle: The functions of utterance-final *then* in spoken English. *Journal of Pragmatics, 43*(14), 3603–3623.

Hasselgren, A. (2002). Learner corpora and language testing: Smallwords as markers of learner fluency. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 143–173). Amsterdam: John Benjamins.

Lenk, U. (1998). *Marking discourse coherence: Functions of discourse markers in spoken English*. Tübingen: Narr.

Mukherjee, J. (2009). The grammar of conversation in advanced spoken learner English: Learner corpus data and language-pedagogical implications. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 203–230). Amsterdam: John Benjamins.

Müller, S. (2005). *Discourse markers in native and non-native English discourse*. Amsterdam: John Benjamins.

Pawley, A., & Syder, F. H. (2000). The one-clause-at-a-time hypothesis. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 163–199). Ann Arbor, MI: The University of Michigan Press.

Peltonen, P. (2018). Exploring connections between first and second language fluency: A mixed methods approach. *The Modern Language Journal, 102*(4), 676–692.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Römer, U. (2005). *Progressives, patterns, pedagogy: A corpus-driven approach to English progressive forms, functions, contexts and didactics*. Amsterdam: John Benjamins.

Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.

Shriberg, E. (1994). *Preliminaries to a theory of speech disfluencies*. PhD dissertation, University of California at Berkeley, CA.

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics, 30*(4), 485–496.

Tottie, G. (2011). *Uh* and *Um* as sociolinguistic markers in British English. *International Journal of Corpus Linguistics, 16*(2), 173–197.

Tottie, G. (2015). *Uh* and *um* in British and American English: Are they words? Evidence from co-occurrence with pauses. In N. Dion, A. Lapierre, & R. Torres Cacoullos (Eds.), *Linguistic variation: Confronting fact and theory* (pp. 38–54). New York: Routledge.