TEXT & TALK SUBMISSION

Full title: Identifying and describing functional discourse units in the BNC Spoken 2014
Short title: Discourse units in the BNC Spoken 2014
Words: 8215
Characters: 44,984

This submission is part of the Special Issue, Corpus Linguistics across the Generations: In Memory of Geoffrey Leech

*Jesse Egbert, Northern Arizona University, jesse.egbert@nau.edu
Box 6032
English Department
Northern Arizona University
Flagstaff, AZ 86001

Stacey Wizner, Northern Arizona University, saw264@nau.edu
Daniel Keller, Northern Arizona University, daniel.ryan.keller@gmail.com
Douglas Biber, Northern Arizona University, douglas.biber@nau.edu
Tony McEnery, Lancaster University and Xi'an Jiaotong University, a.mcenery@lancaster.ac.uk
Paul Baker, Lancaster University, j.p.baker@lancaster.ac.uk

*Main author

Biographies

Identifying and describing functional discourse units in the BNC Spoken 2014

Jesse Egbert, Northern Arizona University
Stacey Wizner, Northern Arizona University
Daniel Keller, Northern Arizona University
Douglas Biber, Northern Arizona University
Tony McEnery, Lancaster University
Paul Baker, Lancaster University

Abstract

On the surface, it appears that conversational language is produced in a stream of spoken utterances. In reality conversation is composed of contiguous units that are characterized by coherent communicative purposes. A large number of important research questions about the nature of conversational discourse could be addressed if researchers could investigate linguistic variation across functional discourse units. To date, however, no corpus of conversational language has been annotated according to functional units, and there are no existing methods for carrying out this type of annotation. We introduce a new method for segmenting transcribed conversation files into discourse units and characterizing those units based on their communicative purposes. The development and piloting of this method is described in detail and the final framework is presented. We conclude with a discussion of an ongoing project where we are applying this coding framework to the British National Corpus Spoken 2014.

Keywords: discourse unit, function, communicative purpose, communicative goal, register, segmentation, speech, BNC Spoken 2014

1. Introduction[1]

Existing corpora of conversational discourse are composed of transcriptions of the language produced by two or more interlocutors during the course of a recording session. These corpora have been invaluable resources for research into the linguistic characteristics of interactive speech, particularly when those research questions focus on (1) general differences between conversation and other registers, or (2) linguistic variation across demographic variables (i.e. social characteristics of the speakers). The largest contemporary corpus of conversational English is the British National Corpus Spoken 2014 (BNC-S 2014). The BNC-S 2014 is composed of 1.5 million words, from 1,251 conversations, that is annotated for many social variables, including the age, gender, geographic dialect, and socioeconomic status of its speakers (Love et al. 2017). This rich annotation makes it possible to answer a wide variety of sociolinguistic research questions where individual speakers are treated as observational units (see, e.g., Brezina, Love and Aijmer 2018; McEnery, Love and Brezina 2017).

Some researchers are interested not in the social characteristics of speakers in the corpus, but rather in the functional and linguistic characteristics of the texts. For the purposes of this study, we use the term *functional* to refer to linguistic features and units that "both perform discourse tasks and reflect aspects of the communicative situation and production circumstances" (Biber 1995: 137). The following remark from Leech (2014: 137) is instructive:

> From the formal point of view, we look at the three main coding levels of linguistic analysis: graphological/phonological, lexigrammatical and semantic. From the functional point of view, we interpret each of the formal levels in terms of three functional tiers: constructing a text (textual function), conveying a representation of some reality (ideational function) and communicating a discourse (interpersonal function). (Leech 2014: 137)

Without 'functional tiers' we are left without the context that is necessary to fully interpret the patterns observed in the 'linguistic analysis'.

Register-based research on functionally-motivated linguistic variation in the BNC-S 2014 is a highly desirable goal (see, e.g., Love et al. 2019), and based on a close examination of individual conversation files in the BNC-S 2014, we hypothesized that there were multiple distinct functional units per file. However, this corpus, like all other corpora of conversation we are aware of, makes it impossible to easily describe the functional characteristics of conversation because the entire speech files do not represent coherent functional units[2] of language. For a corpus of conversation to be useful for

[1] The authors would like to acknowledge the support of Trinity College London, the UK's Economic and Social Research Council (grant ES/R008906/1), the Leverhulme Trust (grant RF-2019-083) and the Northern Arizona University's Corpus Linguistics Research Lab for the work presented in this paper. We would also like to express our gratitude to the coders at Lancaster University, especially Frazer Heritage, Gillian Smith, and Abi Hawtin, who have spent countless hours applying the framework developed here to files in the BNC-S 2014.

[2] We use the term 'functional unit' as a general term for any unit of conversation that is characterized by its discourse function. Below we adopt the term 'discourse unit' to refer to our particular operationalization of functional unit in this study.

addressing register-related research questions, individual files must be segmented into functional units, allowing researchers to investigate the extent to which functional units of conversation differ in their linguistic characteristics. This type of segmentation and annotation has rarely been attempted with a conversational corpus before, so we decided to use the BNC-S 2014 to do this. It is an ideal corpus for our purpose as it is large, varied, recent, publicly available, and richly annotated for speaker meta-data, meaning we can combine the metadata with the annotation of communicative purpose at the level of coherent segments of conversation to make this corpus a powerful resource for answering research questions in areas such as sociolinguistics, discourse analysis, conversation analysis, and register variation. For example, Biber, Egbert, Keller & Wizner (2021) recently used this corpus to produce a taxonomy of conversational discourse types. They showed that conversational talk in the BNC-S 2014 can be segmented into discourse units (henceforth DUs) that have distinct purposes, revealing 16 distinct conversational discourse types that were associated with different communicative purpose profiles.

A first step toward analyzing register variation in conversational corpora is determining whether there are smaller units of conversation that can be reliably identified. A preliminary analysis of the files in the BNC-S 2014 suggested that each of the conversation files contains multiple distinct DUs that have different situational characteristics and communicative goals, which is also true of the demographic portion in the BNC 1994 *inter alia*. Excerpt 1 is an example of a stretch of conversation that seems to transition between two distinct DUs, marked by the broken line:

Excerpt 1

> [1] Speaker A: I didn't want to say that I had not drank it he was really pleased and every time I've been since and he offers me one I say no you're alright
>
> [2] Speaker B: I wonder how much they charge for that privilege? this is special selection of high quality teas delivered from the best tea plantations in the world
>
> [3] Speaker A: mm well they do something in between because it tastes like wee <laughter> it was really disgusting
>
> [4] Speaker B: <laughter> oh dear
>
> [5] Speaker A: I know I shan't be doing that again
>
> --
>
> [6] Speaker B: so what <NAME> was doing today do you know?
>
> [7] Speaker A: I don't know there was some sort of an event I thought she had two days of workshops and things where they've got to take part management stuff I don't know the party was just last night
>
> [8] Speaker B: that was good then so do you think they raffled off that mini?

> [9] Speaker A: no no I think you had to win it I think it was to do with who was the best no I don't think no it wasn't raffled I don't I didn't ask her

The first 5 utterances are the end of a humorous story being told by Speaker A about tea. Speaker B introduces a new DU segment in utterance [6] by asking about a mutual acquaintance. This leads the speakers to a new goal of figuring out the results of a raffle at a party the previous evening.

This type of transition from one communicative goal to another happens dozens of times during the course of this recorded conversation. We can often distinguish among different types of DUs and explicitly refer to them by name when we talk. The concordance lines in Excerpt 2, taken from the BNC-S 1994, provide some examples of this.

Excerpt 2

> And why wasn't she pleased? Anyway to finish that **story** about stopping and starting <F8P>
>
> I'll tell you a small **story**. At one of our constituency surgeries <J9H>
>
> Have you got one other **story** to tell us about your shop? <JNG>
>
> I thought it was quite a good **joke** myself <KB8>
>
> You do recall our **argument** yesterday. <FMN>
>
> And he and my mother had a **disagreement** about this <FY4>
>
> What was your erm wha-- what was your **opinion** of of these new suggestions? <HMP>
>
> as a follow up to the statement that was made on the evening, on the Sunday evening when you gave your **explanation** <HUT>
>
> You've heard my **explanation** of the fact that it was poorly worded <HUD>

A closer review of the broader context of these examples reveals that in every case the speakers are making reference to units of speech (highlighted in bold typeface) that have distinct functions recognized by both the speaker and the listener.

The importance of functional units within spoken language has been acknowledged by scholars in various fields, including Conversation Analysis (see Wald 1976; Jefferson; 1978; Levinson 1979; Houtkoop and Mazeland 1985; Quasthoff, Heller and Morek 2017), sociolinguistics (see Hymes 1967; Forgas 1979; Bakhtin 1986; Tannen 1993; Gumperz 1995; Goldsmith and Baxter 1996), and register studies (see Van Dijk 1981; Crowdy 1995; Biber, Connor and Upton 2007; Egbert and Schnur 2018). However, to date there is no consensus on how to define and operationalize the constructs of a functional unit of conversation.[3] Moreover, there have been very few attempts to actually apply a scheme for speech segmentation and annotation on a large scale. Crowdy (1995), one of the developers of the BNC 1994, wrote:

> The break point between one conversation and the next has to be a fairly subjective decision. Many conversations do not have well-defined openings or closings. Conversations can be interrupted (by another conversation, or an action of some kind) then resumed a few minutes later, or may never be finished. Participants in a conversation may depart, and others may join. Participants may move from one setting to another, still continuing with the same conversation. Changes of topic can fluctuate considerably within the same conversation, or can mark the beginning of a new conversation (p.227).

While not all conversation analysts are likely to agree with all of his assessments, this quote serves to lay out some daunting challenges associated with segmenting and annotating functional units of conversation.

Conversational interactions are a means of accomplishing communicative goals. During the course of a single conversation speakers may cooperatively accomplish multiple communicative goals. Even in cases where there are no interruptions or changes in participants or setting, interlocutors may move from one communicative goal to another many times in a single conversation. For example, during the course of file 'SMHY' in the BNC-S 2014, two friends comment on snow that is falling outside, discuss what gift to purchase for their mothers on Mother's Day, tell stories about stag and hen parties they have attended, share opinions about what makes a good stag or hen party, and make plans for their holiday vacations. Each of these segments of the larger conversation constitutes a distinct unit that could be operationalized as a text. These texts could then be meaningfully described in terms of the functions used to accomplish the goal and the linguistic choices made to achieve those functions. If these descriptive goals are worth working toward—and we believe they are—then we need to analyze conversational discourse at the level of these segments.

To date, however, there have been very few attempts to segment interactive conversations into units that meet the criteria of being natural, self-contained, and functional. These attempts have been limited in scope and success. To our knowledge, a study by Quasthoff, Heller and Morek (2017) is the only large-scale attempt to segment and describe multi-unit turns in conversation from a conversation analytic perspective. They operationalize DU as "the entire conversational unit in which a narrative (or joke, explanation, or other purpose) is interactively prepared, produced and closed" (p.87), thereby identifying three different major genres that DUs can be classified into: narration, explanation, and argumentation. They analyzed an unspecified number of cases in order to

[3] Other frameworks have been developed for segmenting and annotating other discourse domains. Sinclair and Coulthard (1975) propose a well-developed framework for analyzing units of classroom discourse. Their framework included units at five hierarchical levels: lessons, transactions, exchanges, moves, and acts. Their framework has limited applicability to the present study, because it was developed specifically for the analysis of classroom teaching. Indeed, Sinclair and Coulthard were pessimistic about the possibility of analyzing everyday conversations in similar terms, because it is 'the least overtly rule-governed form of spoken discourse' (1975.4). A similar limitation applies to Swales' (1981; 1990) framework for analyzing discourse at the level of rhetorical moves, which is applied almost entirely to written (usually academic) genres.

extract examples of DUs in the three genres. These were then qualitatively analyzed and described for their internal structure and selected linguistic characteristics. Unfortunately, Quasthoff et al. (2017) do not provide enough details about their methods to allow for a replication or an application of their framework. Moreover, for the present study we are interested in defining communicative goal at a more fine-grained level (i.e. *within* a single conversation) than the (macro-)genres of narration, explanation, and argumentation that are the focus of their study. For the most part this goal has been ignored, either because scholars have overlooked its importance or because of the messy nature of conversational language and the relatively subjective nature of the task. Prior attempts suggest these difficulties are not easily dealt with (see, e.g., Crowdy 1995; Biber, Connor and Upton 2007), yet they must be addressed if our goal is to truly understand the situational and linguistic characteristics of conversational language which, predates, presumably, all written registers.

There is no existing method for segmenting natural conversations into smaller, functionally coherent and recognizably self-contained, units. Accordingly, we have developed a methodological framework for segmenting conversations into functional units and annotating their function. In this paper, we introduce the development of our new method for (1) identifying meaningful units of conversational discourse and (2) describing the communicative purposes of these units. In section 2 we introduce the development of our proposed solution and the results of many rounds of piloting the instrument to actual conversational transcripts. In Section 3 we conclude with a discussion of the lessons learned to date, the current status of a project aimed at applying our method to the BNC-S 2014, and our plans for the corpus moving forward.

2. Identifying and describing functional units of conversation: A new framework

Although our method for identifying and describing functional units of conversation evolved considerably as a result of extensive piloting, evaluation, and revisions during the course of many months, two goals remained constant: (1) segmenting and (2) characterizing. Segmenting is the process of identifying boundaries for functional units in the transcripts. Characterizing is the process of describing the segments for their functional attributes. In hindsight, the process of developing our methodological framework can be divided into two major phases. The first and second phases are described in Sections 2.1 and 2.2, respectively. While the methods used to accomplish the first goal of segmenting evolved somewhat during the course of the study, the main difference between the two phases was in the methods used to accomplish the goal of characterizing.

In Phase I, after a *speech event* was identified through segmentation, the coders attempted to assign the unit to a single communicative purpose category. This method provided important insights such as the important communicative purposes that were present in conversational discourse. However, we ultimately determined that this categorical approach to characterization was limited in its ability to account for the complexity inherent in conversation. Therefore, we took what we had learned from this phase and began Phase II where we allowed for the possibility of multiple communicative purposes, coded on an ordinal scale depending on their prominence in the DU. This allowed us to better account for the observed complexity. The next section briefly describes the details of our methods for Phase I before turning to a more detailed description of our final framework in Section 2.2.

2.1. Phase I – Speech events and categorical functions

Our inspiration for the segmentation coding framework came from the Hymesian notion of a speech event; we began by assuming that conversations could be segmented into speech events that existed within speech situations, above the level of speech acts and utterances. Speech events are "activities, or aspects of activities, that are directly governed by rules for the use of speech" (Hymes 1967: 19; see also Hymes 1972; 1974). Hymes never provided an operational definition for a speech event, so we turned to other sources for that. We used aspects of Gumperz's (2015) definition of a speech event as a starting point and began by assuming that each speech event will constitute a 'text', hence we incorporated Egbert and Schnur's (2018) definition of a text into our operational definition of speech event:

1. Recognizably self-contained
2. Sequentially bounded with detectable beginnings and ends
3. Thematically and functionally coherent

We used these three criteria and attempted to segment conversational files into speech event units.

Three of the study's authors attempted to independently segment several texts and then discussed agreement in their placement of speech event boundaries. This was moderately successful as there were some instances where unanimous agreement was achieved by all raters on the boundaries of speech events, and many instances where majority agreement was achieved. While the independent coding was useful, we discovered early on that multiple coders working through a conversation file together was also an effective way to make progress towards a framework. One source of disagreement among raters resulted from differing decisions about the most appropriate level of granularity for speech events, or whether to 'lump' or 'split', when in doubt. To address this, we established a minimum length of four utterances for speech events (later changed to five utterances). While the minimum threshold was inevitably arbitrary, it seemed to work well based on several rounds of coding. When a choice had to be made regarding whether to split a string of utterances into two speech events or lump them into a single speech event, we opted to favor splitting rather than lumping in an effort to fully identify and describe boundaries between functional units.

Subsequent attempts to identify speech event boundaries, both independently and in pairs, showed improvement in our agreement. To further improve reliability, we added an explicit guideline that a speech event boundary must coincide with a shift in communicative goal, not simply with a topic shift. This further improved our agreement. We quantified inter-rater reliability at this point using Pearson's correlations, which we averaged across the purpose categories, to achieve $r = .49$.

At this stage, we had two primary goals: (1) attempt to identify recognizable shifts in communicative purpose to identify speech event boundaries, and (2) attempt to determine the primary communicative purpose for each speech event. Initially we did not have a taxonomy of communicative purposes to choose from. Rather, we were keeping notes on observed communicative purposes that could later be used for classification. Based on these observations, we compiled a list of observed communicative purposes to serve as a starting point for the functional annotation. We benefited from the speech event categories in the taxonomy of speech event functions proposed by Goldsmith and Baxter (1996), though many of the categories in their framework were defined at a more granular level than we were interested in, so we only included communicative purposes that we observed in the coded BNC files. For example, they had separate categories for 'morning talk' and 'bedtime talk'. They also had categories that were defined topically, not functionally, such as 'class information talk' and 'current events talk'.

We began with 12 communicative purposes, which changed many times before the final version. From this point forward, rounds of independent coding included two major steps: segmenting the files into speech events and characterizing those speech events by assigning a communicative purpose category to the segment. It is important to note here that in this phase we only allowed for one communicative purpose per speech event, coded dichotomously (present or absent). We also developed a scheme for XML markup that we used to annotate corpus files for segment boundaries and communicative purpose code. We also introduced the option of coding a file segment as a *non-functional speech event* to annotate sequences of utterances that did not satisfy the three criteria for a speech event.

Further coding led us to modify, based on our experience with the data, the definitional criteria for a speech event to:

1. functionally coherent: A speech event is a sequence of utterances characterized by a single dominant communicative goal.
2. sequentially bounded: A speech event has an identifiable beginning and end.
3. length requirement: A speech event must be a minimum of five utterances or 100 words.

Based on these changes, we carried out further rounds of coding. After each round of coding, the coders would meet to discuss and review differences in segmentation and annotation decisions. Coding was done using an XML markup scheme that required coders to add opening and closing tags for every speech event, with a single communicative purpose code added to each opening tag.

We arrived at a set of ten communicative purposes,[4] with abbreviated codes, not including categories for 'unknown' and 'non-functional speech event'. All functions were informed by observations of the distinctive functions in the conversations themselves. The segmentation agreement among coders was quite high, with our segment boundaries falling within 1-3 utterances the majority of the time, but we found it extremely difficult to reliably assign speech event segments into a single communicative purpose category. One reason for this is that it often seemed plausible there could be more than one communicative purpose for a single speech event, leading us to seriously reconsider our approach and enter into the second phase of our framework development.

2.2. Discourse units and continuous purposes

[4] These were: conversation management, events in progress, joking around, conflict, deliberation, feelings and opinions, intention, advice/suggestion, storytelling, and information.

Phase I established an operational definition for speech event, tested methods for reliably identifying speech event boundaries, and developed a taxonomy of communicative purposes that were observed in conversations in the BNC-S 2014. As a result of our observations during Phase I, we decided to make major adjustments to refine and improve our methodological framework for the segmentation and characterization. With regard to segmentation, we began to hypothesize that the construct of 'speech event', as it was used in previous literature, was inappropriate for the functional units we were observing in conversational discourse. Based solely on the writings of Hymes and Gumperz, it is unclear what the differences are between 'speech events' and 'speech acts'. Moreover, most researchers adopting these terms have done so with the goal of establishing speech event types, where each of the types has a single purpose or function (e.g. Gilner 2016; Friginal et al. 2017). So we made the decision to abandon the term *speech event* as it was apparent we were now doing something quite different than others using that term, adopting instead the term *Discourse Unit* (*DU)* to refer to the focus of this study, i.e., functional segments of conversation.

With regard to the characterization, our observations during Phase I also raised questions about the validity of characterizing functional units using a single communicative function category. So we revisited recent research related to hybrid texts (e.g. Biber, Egbert and Davies 2015) and continuous situational parameters (e.g. Biber and Egbert 2018; Biber, Egbert and Keller 2020). Based on our experience of coding segments in Phase I, we considered the possibility that functional units could be better described using multiple communicative purposes coded continuously on an ordinal scale. We believed that this may help with agreement where coders identified different, yet plausible, communicative purposes. Relatedly, we considered the possibility that these communicative purposes would not all play equal roles in accomplishing the overarching goal of the DU. Thus the ordinal scale represents the degree to which each communicative purpose is used to accomplish the overarching communicative goal of the segment.

Based on these revisions to our framework, we developed a modified set of parameters for DU. A DU is a sequence of utterances that is:

1. Coherent for its overarching **communicative goal**, which is both the primary objective of a DU and the task that the interlocutors are *doing* with language in the DU. This goal is typically coupled with a single topic or theme. Each DU has one communicative goal (e.g. complaining about annoying co-workers; making plans for buying Christmas gifts). There is an open-ended set of specific communicative goals, and these *are not* coded or labeled in our framework.
2. Characterized by one or more **communicative purposes**, where a communicative purpose is a finite set of actions that serve to help accomplish the communicative goal of a DU. Communicative purposes *are* coded in our framework. A DU may rely on one or more communicative purposes. When present, communicative purposes are coded on a scale from 1 – 3.
3. Recognizably self-contained: A DU has an identifiable beginning and end.
4. Length requirement: A DU has a minimum of five utterances or 100 words.

We began piloting this new method, relying on the set of observed communicative purposes developed in Phase I but allowing for multiple functions to be coded on a single DU. Initially, we used a 6-point scale, based on Biber, Egbert and Keller (2020). However, after the first round of pilot coding we determined that six points was too many to make the necessary distinctions and achieve reliable results. Therefore, we adjusted to a 4-point scale for each of the communicative purposes, where:

0 = not present
1 = minor function
2 = major function
3 = dominant function

The primary difference between a score of '2' (major function) and '3' (dominant function) is that a '3' could only be used with one communicative purpose per DU, designating the purpose with a score of '3' as the purpose with the most important role in accomplishing the communicative goal. Not every DU was required to have a dominant purpose. Regardless of whether a DU had a 'dominant purpose' or not, there was no limit to the number of communicative purposes that could be coded with a '1' or a '2' so long as they were functioning to help accomplish the overarching communicative goal of the DU.

This new method was applied in two major steps: (1) Segmentation: identify the boundaries of a DU, and (2) Characterization: code each communicative purpose for the degree to which it is actively helping to accomplish the overarching communicative goal of the DU (0 – 3).

Excerpt 3 below is an example from the BNC-S 2014 of a DU with an overarching goal of learning about each other's children. This includes multiple purposes that are present to varying degrees. These purposes include figuring out the exact age of Speaker B's toddler, describing his fascination with a vacuum cleaner, and a narrative from Speaker A about similar behavior from their child in the past. Although these purposes are distinct from each other, they converge in this segment to create a single DU that is coherent for its overarching communicative goal.

Excerpt 3

fto="1" des="2" pas="2"
Speaker A: so how old is your kid?
Speaker B: well his uh a twenty-one month old so
Speaker A: oh yeah
Speaker B: he's quite small <pause> yeah he's gonna be two in June
Speaker A: so the hoover is probably still quite interesting
Speaker B: oh yeah he's fascinated by it <pause> absolutely <pause> yeah
Speaker A: fascinated
Speaker B: just every time I hoover he just you know he just wants to grab it and just do it all himself <pause> maybe I should buy him a toy one or something
Speaker A: I remember that <pause> I remember my son used to <pause> he was so interested in the hoover <pause> he also used to pull it out all the time pull it out

> of the plug and um and then it’s kind of changed you know the hoover goes on it's like shut up I want to watch the TV I want to watch the TV

Feeling satisfied with our framework as refined in Phase II, we attempted a systematic evaluation of inter-rater reliability for files coded by four of the study’s authors. This presented several challenges. Methods for inter-rater reliability assume that the objects being rated have fixed boundaries. As disagreements on DU boundaries rendered it impossible to directly compare sequences of utterances, at this stage we did not evaluate the reliability of segmentation boundaries. However, we did measure inter-rater agreement in communicative purpose categorization at the utterance level, simultaneously accounting for boundaries and classification reliability. While this did not directly provide information about the two constructs of segmentation and characterization separately, it did allow us to measure inter-rater agreement across all of the communicative purposes, and for each of them separately.

Inter-rater agreement was measured using Krippendorf’s alpha and Pearson’s correlations. Krippendorf’s alpha measured agreement across all four raters in dichotomous terms: presence or absence of a code for each communicative purpose. Pearson’s correlations measured agreement between every possible pair of raters in continuous terms: the degree of association in scores (0-3) for each communicative purpose. Correlation values were averaged across all of the possible pairs of raters for each of the communicative purposes. While these measures did not directly account for DU boundaries, taken together they did provide some insight into agreement among coders.

In our first pilot, we achieved an overall Krippendorf’s alpha value of .29, a “fair agreement” (Landis and Koch (1977), and a moderate mean Pearson’s correlation of .49. We also noted, though, that there was extreme variability in inter-rater agreement across communicative purposes, leading us to make refinements to the operational definitions for several purposes. A second round of pilot coding resulted in modest improvements to both Krippendorf’s alpha (.33) and Pearson’s correlation (.50) and much less variability across communicative purposes. For both pilot rounds, we noted that some of the communicative purpose categories were much more common in some files than others. Thus, the agreement estimates for the less frequent purposes were less robust.

We combined the quantitative measures of agreement with a qualitative, manual review of agreement across files. This was done using a vertically aligned spreadsheet for each file that had one column for each rater, one row for each utterance, and the communicative purpose code assigned by that rater for that utterance. This allowed us to evaluate (mis)alignment in DU boundaries and (dis)agreement in communicative purpose codes, helping us to make sense of the quantitative reliability results and determine systematic areas of disagreement. Crucially, it revealed some important trends that could not be observed from the quantitative measures. First, minor disagreements in DU boundaries had a strong negative effect on both quantitative measures of agreement, even though boundaries were not accounted for in any direct way. This was mostly due to entire DUs being assigned the same set of communicative purposes. Second, the difference between a communicative purpose code of 0 and 1 had a much stronger effect on the alpha statistic than a difference between 1 and 2 or 2 and 3 because this measure only accounted for presence or absence, not the full ordinal scale. Third, and most importantly, there were many cases where raters disagreed, yet after closer inspection and discussion we determined that both coding decisions were equally plausible.

This final observation led us to the conclusion that quantitative measures of agreement or reliability were not accurately reflecting the nature of the task as they assume that there is one objectively correct answer that raters are attempting to converge on. Our observations up to this point in the project suggest that (1) there is considerable evidence that there is an objective reality to functional DUs, but (2) there is a permissible degree of flux that exists in the boundaries between DUs and the degree to which communicative purposes are present within those units. The constructs in our study, and the way we have operationalized them—DU boundaries, presence of communicative purposes, extent to which purposes are helping to achieve the overarching communicative goal of the unit—are all, to some extent, dependent on the observer's perception of the nature of the interaction and the intents of the speakers involved. Hence, we decided that the ultimate goal was *not* perfect agreement among independent coders. Rather, our aim was for independent coders, trained using the same framework, to make coding decisions that other trained coders would deem plausible. So our goal is for raters to achieve segmentation boundaries and communicative purpose ratings that would be deemed plausible by another trained rater. Thus, plausibility checks are the basis for feedback to raters-in-training and for evaluating the degree to which segmentation and characterization has been successful. Moving forward, however, we also plan to conduct large-scale evaluations of inter-rater agreement. This will be done using the methods presented in Section 2.2, as well as other methods we are currently exploring that will allow us to account for agreement in (1) segmentation boundaries and (2) the dichotomous (presence/absence) of communicative purposes.

Additional rounds of pilot coding were carried out to test our new plausibility criterion and to make further refinements to the coding framework. We are quite satisfied with the plausibility criterion. It can be usefully applied at all three steps of the coding process: (1) plausibility of DU boundaries, (2) plausibility of communicative purpose presence, and (3) plausibility of communicative purpose degree. Checks at these three stages have been an effective means of providing feedback to new coders during the training process and are a useful method for more experienced coders to periodically calibrate their coding with others to ensure that there is no drift in their understanding and application of the coding framework and communicative purposes.

The final stages of pilot coding resulted in small changes to the parameters of communicative purpose. We also began annotating where both recording-related talk (language that refers to the task of recording the conversation) and foreign language (any language other than English). These do not take the place of communicative purposes, rather they are added as XML markup to the unit where the recording-related talk or foreign language were observed. These annotations can be used in both DUs and non-functional DUs. As noted, in the early stages of developing our coding framework we observed that most, but not all, normal conversation is structured and organized as functional units. We consistently observed stretches of conversation that did not function as a coherent DU. This is not merely an artifact of our length requirement. While non-functional DUs are often small, they can also be more extended. We believe there are important theoretical implications of both the discoveries that (1) most utterances in a conversation can be organized into larger units of conversation and (2) this is not always the case and there can be portions of conversation that do not function together as larger units. An example of a non-functional DU can be seen in Supplement C.

2.3. Final coding framework

We now describe the final coding framework, introduce and illustrate the communicative purposes, and provide an example of conversation coded using the framework. Supplements A and B represent the sum total of our work over many months to develop a comprehensive method for identifying and describing functional units of conversational discourse. Our final coding framework and instructions are included in Supplement A. This document contains the full set of guidelines that coders use to (1) segment transcribed speech files into DUs, and (2) characterize those DUs according to their communicative purposes. These guidelines are divided into three sections. Section one provides operational definitions for DUs, communicative goals, and communicative purposes, and contains a listing of the set of communicative purposes in the framework. Section two contains the full set of steps that coders follow when segmenting and characterizing DUs. Section three establishes the XML markup coders use when annotating the corpus files.

The full set of communicative purposes are contained in Supplement B below. This contains labels and definitions for each of the communicative purposes in the framework. We will briefly describe and illustrate each of these nine communicative purposes here. It should be noted that the examples are meant to demonstrate when a particular purpose would be present; the degree to which the purpose is present, as well as the possible presence of additional purposes, varies from example to example.

1. **Situation-dependent commentary (sdc).** Occurs when speakers in a conversation are commenting on people or objects that are present, or events that are occurring in their shared situational context. Examples include (1) commentary on the unsafe driving practices of another driver at the petrol station where they are waiting for their turn at the pump, and (2) conversation about rules and strategies in a board game as it is being played.

2. **Joking around (jok).** Conversation that is intended to be humorous, including both lighthearted and darker humor. It also includes good humored banter, teasing and flirting. Examples include (1) a hyperbolic comparison between a bad tasting pie and sawdust, and (2) one speaker teasing another because her blouse was being worn inside out.

3. **Engaging in conflict (con).** Includes disagreement of any type, including lighthearted debate as well as more serious quarreling. Examples include (1) a debate over which key on a key ring fits which door in a house, and (2) a friendly disagreement over which of the two speakers is more likely to become rich one day.

4. **Figuring things out (fto).** Discussion aimed at exploring or considering options or plans, including discussion about how things work and what the best solution to a

problem may be. Examples include (1) discussion about the appropriateness of visiting in-laws after a spouse's death, and (2) attempts to understand and explain the recent behavior of a mutual acquaintance.

5. **Sharing feelings and opinions (fel).** Discussion about feelings, opinions, and beliefs, including the airing of grievances and the sharing of personal perspectives. Examples include (1) an explanation to justify a speaker's preference for an item of clothing, and (2) a discussion about political views.

6. **Giving advice and instructions (adv).** Occurs when one speaker offers directions, advice, or suggestions to another speaker. Examples include (1) one speaker helping another to navigate a website to order tickets by giving step-by-step commands during the process, and (2) one speaker offering suggestions to another on the best kind of copy paper to purchase

7. **Describing or explaining the past (pas).** Narrative stories about true events from the past or other references to people or events from the past. Examples include (1) a speaker telling stories from a favorite vacation, and (2) two speakers reminiscing together about the past while sorting through boxes stored in the attic.

8. **Describing or explaining the future (fut).** Descriptions or speculations about future events and intentions, including those that are planned and those that are more hypothetical. Examples include (1) one speaker describing plans for a date with a significant other, and (2) two speakers sharing their plans for life after graduation from university.

9. **Describing or explaining (time-neutral) (des).** Descriptions or explanations about facts, information, people or events where time (past or future) is either irrelevant or unspecified. Examples include (1) a speaker responding to another's questions about the progress of house renovations, and (2) a description of the difference between two products.

Each coder in the study participated in an extensive training process. During this process, one or more of the project leaders described the broad aims of the study and the design of the BNC. Coders were then introduced to the coding framework and the list of communicative purposes, and saw multiple examples of speech files, both uncoded and coded. Each coder then performed multiple rounds of coding. After each of these rounds, project leaders reviewed the coding for plausibility and offered feedback during follow-up meetings. Once a coder's work was consistently achieving high levels of plausibility, they were assigned a batch of files to code independently.

Excerpt 4 below is an example of a sequence of two DUs and a non-functional DU. For readability, all XML markup has been excluded. The same segment is included in Supplement C with all XML markup.

Excerpt 4

Discourse unit 1: sdc="3" fto="2" jok="2"

Speaker A: yeah he looks happy
Speaker B: yeah he looks happy then you've got the cow hanging up who's skinned and then you've got a piece of steak
Speaker A: but it doesn't look like they're gonna be selling meat does it?
Speaker B: no
Speaker A: so what is it?
Speaker B: what is it?
Speaker A: I thought it was an internet café
Speaker B: but it looks like it's selling meat from the counter
Speaker A: maybe it's a a meat themed internet café
Speaker B: yeah look they've got cows inside too
Speaker A: oh

Non-functional discourse unit

Speaker B: oh that's interesting maybe they're er they're still setting up
Speaker A: well you know what it's like round here it'll probably stay like that for two years and then just disappear

----

Discourse unit 2: pas="3" fel="2" jok="1"

Speaker B: so <NAME> learnt to wave? she waved today
Speaker A: yeah she waved at me this morning
Speaker B: oh she waved at me at lunchtime <laughter>
Speaker B: oh it was so cute <laughter>
Speaker B: she didn't eat much at lunchtime I went home and she was in asleep
Speaker A: yeah

In DU 1 two speakers are commenting on an advertisement they see in a shop for a cow that is for sale. In addition to the dominant purpose of ‘situation-dependent commentary’, this DU also includes the purposes of ‘figuring things out’ and ‘joking around’. The speakers then briefly give their attention to commenting on something different, but this was coded as a ‘non-functional DU’ because it did not meet the minimum length requirement and it does not function as part of either of the two adjacent DUs. In DU 2, the speakers each share similar experiences from earlier that day when a mutual acquaintance, presumably a young child, waved at them. This was coded with the dominant purpose of ‘describing or explaining the past’, along with the major purpose of ‘sharing feelings or opinions’ and a minor purpose of ‘joking around’.

3. Conclusions, and beginnings

Conversational speech is a rich discourse domain that can be segmented into functionally coherent DUs. These can be characterized for one or more communicative purposes that serve to help speakers achieve larger communicative goals. The coding framework introduced in Section 2, accompanied by the Supplements, represents our proposed framework for segmenting and characterizing DUs in English. We believe this novel framework for identifying and describing DUs and their communicative purposes could be usefully applied to answer a vast array of important research questions that have never been tackled before. We hope that other researchers will find uses for the framework, and the coded BNC-S 2014 files we are currently applying it to. We also sincerely hope that this framework and its application thus far acts as a springboard to further attempts to develop and refine methods for segmenting and characterizing conversations. It would be encouraging to see future research that applies and adapts this framework for specific registers of interactive spoken discourse (e.g. interviews, business meetings), as well as to languages other than English.

We currently have two teams of coders, one at Lancaster University and another at Northern Arizona University, who are coding a subset of the files from the BNC-S 2014 corpus. This subset includes a 50% sample (*n* = 479) of the 958 files containing two or three speakers. Our framework was developed for and piloted on files with only 2-3 speakers. In order to select this subset we first rank ordered the files by number of utterances and sampled the middle 50%. Thus, the shortest 25% and longest 25% were excluded from the sample. At some point in the future we hope to have the resources to complete the coding of all files in the BNC-S 2014. The coded dataset will eventually be made available to download for use by other researchers. We hope future research will explore the extent to which this can be applied to conversations with four or more speakers.

Once these files are coded, we will begin a series of projects to address a range of interesting linguistic research questions. This data will allow researchers to explore new insights into questions that have not been adequately answered in any previous study we are aware of. In one sense, this coding framework and the coded portion of the BNC-S 2014 corpus open up an entirely new sub-field of descriptive corpus linguistics, capable of addressing questions about linguistic variation across functional units of conversation, defined by communicative goals shared between interlocutors. We will be able to explore how this language variation interacts with variation across demographic characteristics of the speakers and the internal structure of DUs that have dominant purposes (such as ‘figuring things out’, ‘sharing feelings and opinions’, and ‘describing or explaining the past’) based on a generalizable sample of thousands of DUs *inter alia*. We are excited by these possibilities and look forward to seeing how this data will be used, how this framework will be applied to other conversational texts and corpora, and how it will be built upon in future research related to functional units of conversation that can be used to “interpret each of the formal levels” of linguistic analysis (Leech 2014: 137).

References

Bakhtin, Mikhail Mikhailovich. 1986. The problem of speech genres. In Caryl Emerson & Michael Holquist (eds.), *M.M. Bakhtin: Speech genres and other late essays*,60-102. Austin: University of Texas Press.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison.* Cambridge: Cambridge University Press.

Biber, Douglas, Ulla Connor & Thomas Upton. 2007. *Discourse on the move: Using corpus analysis to describe discourse structure* (Vol. 28). Amsterdam: John Benjamins Publishing.

Biber, Douglas & Jesse Egbert. 2018. *Register variation online.* Cambridge: Cambridge University Press.

Biber, Douglas, Jesse Egbert & Mark Davies. 2015. Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora* 10(1). 11-45.

Biber, Douglas, Jesse Egbert & Daniel Keller. 2020. Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory, 16(3), 581-616.*

Biber, Douglas, Jesse Egbert, Daniel Keller, & Stacey Wizner. 2021. Towards a taxonomy of conversational discourse types: An empirical corpus-based analysis. *Journal of Pragmatics, 171,* 20-35.

Brezina, Vaclav, Robbie Love, & Karin Aijmer. (eds.). 2018. *Corpus approaches to contemporary British speech: Sociolinguistic studies of the Spoken BNC2014*. New York: Routledge.

Crowdy, S. 1995. *The BNC spoken corpus* in Geoffrey Leech, Greg Myers & Jenny Thomas, (eds.), *Spoken English on computer: transcription, mark-up and application* Harlow: Longman, pp. 224-235.

Egbert, Jesse, Douglas Biber & Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology* 66(9).1817-1831.

Egbert, Jesse & Erin Schnur. 2018. Missing the trees for the forest: The role of the text in corpus and discourse analysis. In Anna Marchi and Charlotte Taylor (eds.), *Corpus approaches to discourse: A critical review,* New York: Routledge.

Forgas, Joseph. 1979. *Social episodes: The study of interaction routines.* New York: Academic Press.

Friginal, Eric, Joseph Lee, Brittany Polat, & Ashley Roberson. 2017. Corpora of spoken Academic discourse and learner talk: A survey. In *Exploring spoken English learner language using corpora* (pp. 35-63). Palgrave Macmillan, Cham.

Gilner, Leah. 2016. Identification of a dominant vocabulary in ELF interactions. *Journal of English as a Lingua Franca*, *5*(1), 27-51.

Goldsmith, Deanna & Leslie Baxter. 1996. Constituting relationships in talk: A taxonomy of speech events and social relationships. *Human Communication Research*, 23(1): 87 – 114.

Gumperz, John J. 2015. Interactional sociolinguistics: A personal perspective. In Deborah Tannen, Heidi Hamilton & Deborah Schiffrin (eds.), *The handbook of discourse analysis*.

Houtkoop, Hanneke and Harrie Mazeland. 1985. Turns and discourse units in everyday conversation. *Journal of Pragmatics* 9: 595–619.

Hymes, Dell. 1967. Models of the interaction of language and social setting. *Journal of Social Issues*, *23*(2), 8-28.

Hymes, Dell. 1972. Models of the interaction of language and social life. In John Gumperz & Dell Hymes (eds.), *Directions in sociolinguistics: The ethnography of communication* (pp.35-71). New York: Holt, Rhinehart & Winston.

Hymes, Dell. 1974. *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press.

Jefferson, Gail. 1978. Sequential aspect of story-telling. In Jim Schenkein (ed.), *Studies in the organization of conversational interaction*. New York: Academic Press, 219–248.

Landis, J. Richard and Gary G. Koch. 1977 The measurement of observer agreement for categorical data. *Biometrics*, 33, pp. 159–174.

Leech, Geoffrey. 2014. *Language in literature: Style and foregrounding*. New York: Routledge.

Levinson, Stephen. 1979. Activity types and language. *Linguistics* 17: 365–399.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina, & Tony McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3), 319-344.

Quasthoff, Uta, Vivien Heller & Miriam Morek. 2017. On the sequential organization and genre-orientation of discourse units in interaction: An analytic framework. *Discourse Studies* 19(1): 84 – 110.

Sinclair, John & Malcolm Coulthard. 197*). Towards an analysis of discourse: The English used by teachers and pupils*. London: Oxford University Press.

Swales, John. 1981. *Aspects of article introductions*. (LSU research reports 1). LSU, Aston University, Birmingham.

Tannen, Deborah. (ed.). (1993). *Framing in discourse.* New York: Oxford University Press.

Swales, John. 1990. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Van Dijk, Teun. 1981. Episodes as units of discourse analysis. In Deborah Tannen (ed.), *Analyzing discourse: Text and talk*. (177 – 195). Georgetown: Georgetown University Press.

Wald, Benji. 1978. Zur Einheitlichkeit und Einleitung von Diskurseinheiten. In Uta Quasthoff (ed.), *Sprachstruktur – Sozialstruktur*. Königstein: Scriptor 128–149.